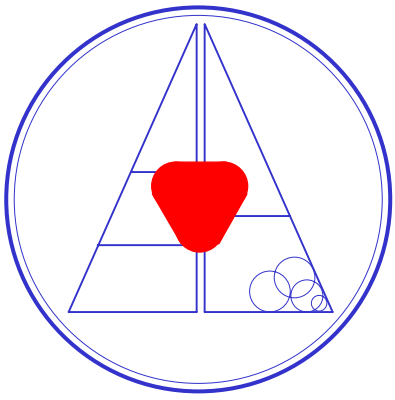
Handbook for

Health Care Research



Robert L. Chatburn, BS, RRT, NP-S, FAARC

Director

Respiratory Care Department

University Hospitals of Cleveland

Associate Professor

Department of Pediatrics

Case Western Reserve University

Cleveland, Ohio

Mandu Press Ltd

Cleveland Heights, Ohio

Published by:
Mandu Press Ltd
PO Box 18284
Cleveland Heights, OH 44118-0284

All rights reserved. This book, or any parts thereof, may not be used or reproduced by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the publisher, except for the inclusion of brief quotations in a review.

First Edition

Copyright © 2003 by Robert L. Chatburn

Library of Congress Control Number: 2003103283

ISBN, printed edition: 0-9729438-0-3

ISBN, PDF edition: 0-9729438-1-1

First printing: January 2003

Sedond printing: February 2004

Care has been taken to confirm the accuracy of the information presented and to describe generally accepted practices. However, the author and publisher are not responsible for errors or omissions or for any consequences from application of the information in this book and make no warranty, express or implied, with respect to the contents of the publication.

Table of Contents

SECTION I INTRODUCTION 1

Chapter 1. Why Study Research? 1
The Importance of Research in Health Care 1
Health Care Education 2
Professional Accountability 3
Administration of Health Care Services 3
Evaluating New Equipment and Methods 4
Questions 5
Definitions 5
True or False 5
Multiple Choice 5

Chapter 2. Ethics and Research 6
Institutional Review and Human Subjects' Rights 6
Functions of the Institutional Review Board 6
Composition of the Institutional Review Board 7
Approval of the Institutional Review Board 7
Informed Consent 8
Background 8
Role Today 9
Revocation of Consent 9
Ethical Issues 10
Basic Principles 10
Objective Patient Care 12
Reporting Research Results 12
Questions 12
Definitions 12
True or False 13
Multiple Choice 13

Chapter 3. Outcomes Research 14
A Brief History 14
Understanding the Jargon 15
Outcomes Research: Focus and Methods 16
The Outcome of Outcomes Research 20
Examples from Respiratory Care 21
Benchmarking 23
Summary 24
Questions 25
Definitions 25
True or False 25
Multiple Choice 25

SECTION II PLANNING THE STUDY 6

Chapter 4. The Scientific Method 27
The Scientific Method 27
1. Formulate a Problem Statement 27
2. Generate a Hypothesis 27
3. Define Rejection Criteria 28
4. Make a Prediction 28
5. Perform the Experiment 28
6. Test the Hypothesis 29
Steps in Conducting Scientific Research 29
Develop the Study Idea 29
Search the Literature 29
Consult an Expert 29
Design the Experiment 29
Write the Protocol 30
Obtain Permission 30
Collect the Data 30
Analyze the Data 30
Publish the Findings 30
Questions 31
Definitions 31
True or False 31
Multiple Choice 31

Chapter 5. Developing the Study Idea 32
Sources of Research Ideas 32
Developing a Problem Statement 33
Judging the Feasibility of the Project 35
Significance of the Problem 35
Measurability of the Problem 35
Time Constraints 36
Availability of Subjects 36
Cost and Equipment 36
Experience 36
Summary 37
Questions 38
Definitions 38
True or False 38
Multiple Choice 38

Chapter 6. Reviewing the Literature 40
Conducting the Literature Serach 40
Scope of the Review 40
Performing the Search 41
Sources of Information 42
Books 42

Journal Articles 42
The Internet 43
How to Read a Research Article 44
The Abstract 44
The Introduction 44
The Methods Section 44
The Results Section 46
The Discussion 47
Conclusion 47
Summary 47
Questions 47
True or False 47
Multiple Choice 48

Chapter 7. Designing the Experiment 49
Samples and Populations 49
Methods of Obtaining a Sample 50
Basic Concepts of Research Design 52
Experimental Designs 53
Pre-Experimental Designs 53
Quasi-Experimental Designs (Case Control) 54
True Experimental Designs (Randomized Control) 55
Analysis of variance (ANOVA) 57
Validity of Research Designs 61
Non-Experimental Study Designs 62
Retrospective Studies 63
Prospective Studies 63
Case Studies 64
Surveys 64
Correlational Studies 64
Questions 65
Definitions 65
True or False 65
Multiple Choice 65

SECTION III CONDUCTING THE STUDY 27

Chapter 8. Steps to Implementation 67
Writing the Study Protocol 67
Creating a General Plan 67
The IRB Study Protocol Outline 68
Funding 73
American Respiratory Care Foundation 73
Data Collection 73
The Laboratory Notebook 74
Specialized Data Collection Forms 78

Computers 78
Questions 79
True or False 79

Chapter 9. Making Measurements 80
Basic Measurement Theory 80
Accuracy 80
Precision 82
Inaccuracy, Bias and Imprecision 82
Linearity 83
Calibration 84
Sources of Bias (Systematic Error) 85
Sources of Imprecision (Random Error) 89
Measuring Specific Variables 90
Pressure 90
Flow 92
Volume 95
Humidity 99
Signal Processing 100
Recording and Display Devices 101
Questions 101
Definitions 101
True or False 102
Multiple Choice 102

SECTION IV ANALYZING THE DATA 105

Chapter 10. Basic Statistical Concepts 105
Preliminary Concepts 105
Definition of Terms 105
Levels of Measurement 106
Significant Figures 108
Zeros as Significant Figures 108
Calculations Using Significant Figures 109
Rounding Off 109
Descriptive Statistics 109
Data Representation 109
Measures of the Typical Value of a Set of Numbers 114
Measures of Dispersion 116
Correlation and Regression 119
Inferential Statistics 122
The Concept of Probability 122
The Normal Distribution and Standard Scores 125
Sampling Distributions 127
Confidence Intervals 130
Error Intervals 132

Data Analysis for Device Evaluation Studies....137
Interpreting Manufacturers' Error Specifications....141
Hypothesis Testing....144
Type I and II Errors....150
Power Analysis and Sample Size....152
Rules of Thumb for Estimating Sample Size....155
Clinical Importance Versus Statistical Significance....160
Matched Versus Unmatched Data....160
Questions....161
Definitions....161
Multiple Choice....162

Chapter 11. Statistics for Nominal Measures....167
Describing the Data....167
Characteristics of a Diagnostic Test....168
True and False Positive Rate....171
True and False Negative Rate....171
Sensitivity and Specificity....171
Positive and Negative Predictive Value....172
Diagnostic Accuracy....172
Likelihood Ratio....172
Receiver Operating Characteristic (ROC) Curve....173
Correlation....174
Kappa....174
Phi....175
Comparing a Single Sample With a Population....176
Binomial Test....176
Z Test....177
Comparing Two Samples, Unmatched Data....178
Fisher Exact Test....178
Comparing Two or More Samples, Matched Data....179
McNemar's Test....179
Comparing Three or More Samples, Unmatched Data....181
Chi-Squared Test....181
Questions....182
Definitions....182
True or False....182
Multiple Choice....182

Chapter 12. Statistics for Ordinal Measures....184
Describing the Data....184
Correlation....184
Spearman Rank Order Correlation....184
Comparing Two Samples, Unmatched Data....186
Mann-Whitney Rank Sum Test....186
Comparing Two Samples, Matched Data....187
Wilcoxon Signed Rank Test....187

Comparing Three or More Samples, Unmatched Data....187
Kruskall-Wallis ANOVA....187
Comparing Three or More Samples, Matched Data....188
Friedman Repeated Measures ANOVA....188
Questions....190
Multiple Choice....190

Chapter 13. Statistics for Continuous Measures....192
Testing for Normality....192
Kolmogorov-Smirnov Test....192
Testing for Equal Variances....193
F Ratio Test....193
Correlation and Regression....194
Pearson Product Moment Correlation Coefficient....195
Simple Linear Regression....196
Multiple Linear Regression....197
Logistic Regression....197
Comparing One Sample to a Known Value....200
One sample *t*-test....200
Comparing Two Samples, Unmatched Data....200
Unpaired *t*-test....200
Comparing Two Samples, Matched Data....202
Paired *t*-test....202
Comparing Three or More Samples, Unmatched Data....204
One Way ANOVA....205
Two Way ANOVA....206
Comparing Three or More Samples, Matched Data....210
One Way Repeated Measures ANOVA....210
Two Way Repeated Measures ANOVA....212
Questions....217
Multiple Choice....217

SECTION V PUBLISHING THE FINDINGS....220

Chapter 14. The Paper....220
Selecting an Appropriate Journal....220
Getting Started....221
The Structure of a Paper....222
Title....222
Abstract....222
Introduction....223
Methods....223
Results....224
Discussion....225
Conclusion....225
Illustrations....225

Submission for Publication 226
First Steps 226
Peer Review 226
Revision 226
Production 226
Mistakes to Avoid 227
Questions 228
True or False 228

Chapter 15. The Abstract 229
Background 229
Specifications 229
Content Elements 229
Format 230
Template 230
Model Abstract 231
Example Template for Submiting an Abstract 231
Model Abstract #1 231
Model Abstract #1 232
Model Abstract #2 233
What not to do (analysis of rejected abstracts) 234
Summary 239
Questions 239
True or False 239

Chapter 16. The Case Report 240
Who Should Write It? 241
Attributes of a Reportable Case 242
A New Disease or Condition 242
A Previously Unreported Feature or Complication 243
A Particularly Instructive Example of a Known Condition 243
A Case Illustrating a New Diagnostic Test or Monitoring Technique 243
A New Treatment Modality 243
A New Outcome of Treatment 243
Steps in Preparing a Case Report 243
Identification of an Appropriate Case 244
Review of the Pertinent Literature 244
Consultation and Discussion 245
Planning the Paper and Assignment of Roles and Authorship 245
Further Investigation of the Case 245
Preparation of the First Draft 246
Preparation of Tables and Illustrations 246
Consultation and Discussion 246
Revision of Manuscript 246
Preparation and Submission of Final Draft 246
Structure of a Case Report 247
Introduction 247

Case Summary 247
Tables and Illustrations 248
Discussion 248
References 249
Avoiding Common Mistakes in Case Report Writing 249
Tunnel Vision 250
Insufficient Documentation of Case 250
Insufficient Documentation of Intervention 251
Poor Patient Care 251
Erroneous Premise 251
Submission to the Wrong Journal 251
Literary Inexperience 251
Inadequate Literature Review 252
Ineffective Illustrations or Tables 252
Poor References 252
Technical Mistakes 252
Failure to Revise the Manuscript after Editorial Review 252
Questions 253
True or False 253

Chapter 17. The Poster Presentation 254
Layout 254
Planning 254
Materials 255
Questions 256
True or False 256

SECTION VI APPENDICES 220

Appendix A. Glossary

Appendix B. Peer Review Checklists

Appendix C. Model Paper

Appendix D. Response to Reviewers

Appendix E. Answers to Questions

Appendix F. Statistics Selector

PREFACE

Learning to do research is like learning to ride a bicycle, reading a book is not much help. You need to learn by doing, with someone holding you up the first few times. Yet, the student of health sciences research must be familiar with basic concepts that can be studied by reading. The trick is to select the right topics and present them in a way that is both relevant and interesting.

This book is the result of over 20 years of experience doing research in the field of respiratory care. I have tried to select topics and statistical procedures that are common in medical research in general, and to allied health care in particular. It is by no means an exhaustive treatise on any particular aspect of medical research. Rather, it is more of a practical guide to supplement specialized statistics textbooks. Yet it can function as a stand-alone text for a short course in research in a 2 or 4 year respiratory care or other allied health program. In fact, this book grew out of the notes I have used for the last 6 years to teach research at Cuyahoga Community College.

At one level, the book is geared for the student or health care professional who wants to become involved with research. Basic concepts are presented along with real world examples. Naturally, because I am a respiratory therapist, the examples have to do with respiratory care. However, the concepts are applicable to any area of medical research. I have tried to keep the theory and mathematics at the most basic level. I assume that the reader will have basic computer skills and will have access to software that will handle the math. For that reason, unlike many books on the topic, there are no probability tables for calculating things like the critical values of the *t* statistic. Computers have made hand calculations all but obsolete. What the student really needs to know is which procedure to use, when, and why.

For the experienced researcher, the book is organized so that basic research procedures and definitions can be quickly looked up. This is important because when you are in the middle of a project you don't want to be interrupted to pour through pages and pages of theory when all you want is to be reminded of which test to use or how to format the data for computer entry.

Not every health care professional will be directly involved with research. However, everyone will be involved with the results of research. And most will be involved at some time with some sort of continuous quality improvement project, which will inevitably require some familiarity with research techniques. Therefore, this book, if nothing else, should serve as an excellent tool to help you become an "educated consumer" of research. After all, how can you appreciate the information in professional journals if you don't even know what a *p* value is? Researchers who publish in journals are trying to sell you their ideas. If you don't understand the procedures they use to generate the ideas and the language they use to sell them, you could end up buying a "lemon".

There are several features in this book that I think are unique. For example, the descriptions of statistical tests are standardized in a practical format. For each procedure, a hypothetical (or sometimes real-world) study problem is introduced, the hypothesis is stated, the data are given in the format that they are entered into the computer, then a detailed report from an actual statistical program is given.

Another unique feature is the chapter on writing the stand-alone abstract. The new researcher's first experience with publishing research will usually be in the form of an abstract, rather than a full text article. For this reason, I have placed particular emphasis on how to write an abstract that will pass peer review. There is a model abstract that has actually been published in Respiratory Care journal along with several abstracts that were submitted but rejected. I review each abstract in detail, just as I did when I reviewed them for the journal, and explain the mistakes made. These detailed examples are intended to

give the reader a sense of having a mentor looking over their shoulder giving help and encouragement. Just like riding a bike. In fact, the text throughout is worded in almost a conversational style. This really helps to illustrate the relevance of each new concept that might otherwise seem dull and intangible.

Also included in the Appendices is a model manuscript that was actually published in Respiratory Care. Not only that, but the comments of the peer reviewers is included along with the authors' responses. One of the biggest obstacles for new researchers is that they have a hard time accepting critical comments about a manuscript they have submitted for publication. Many, maybe even most, are so discouraged that they do not make the suggested revisions and their work goes to waste. My hope is that reading the reviewer's comments and the responses, you will get the idea that (1) every researcher, no matter how experienced, will be criticized and (2) the criticism only leads to a better product if you follow through. I always tell my students that the very first thing they have to learn is to "put your ego on the shelf".

Robert L. Chatburn, RRT, FAARC

Cleveland, Ohio

March, 2002

DEDICATION

Allied health professionals are rarely given formal training in research methodology. And even when they are, it is never more than a cursory overview. The real learning happens in apprenticeship. One must have a good mentor who can pass on the benefit of his knowledge and experience. I have been blessed with three of the best mentors a person could have.

The first is Marvin Lough, MBA, RRT, FAARC. Marv gave me my first job in the profession and helped me create a dedicated research position. He taught me that it is not what a person holds in memory that counts, but rather what he knows how to find. He has exemplified to me, in every way, what it means to be a professional, a leader, and a gentleman.

The second is Frank P. Primiano Jr., PhD. Frank has the most disciplined, logical and penetrating mind that I have ever encountered. He taught me the basic skills of a scientist. He taught me that brilliance lies in paying attention to the details and the supreme importance of defining and understanding the words you use. But most importantly, he taught me "If you explain something so that even a *fool* can understand it…then only a fool *will* understand it."

The third is Terry Volsko, BS, RRT, FAARC. She would say that *I* am *her* mentor, but the truth is that she has taught me as much as I have taught her. I have never met anyone with a greater hunger for knowledge or a stronger will to succeed. She has been a brilliant and tireless student, an insightful critic, and a compassionate friend. My other mentors showed me how to succeed; Terry showed me why.

SECTION I INTRODUCTION

Chapter 1. Why Study Research?

The chances that you, the reader, will become a famous researcher may be slim. For example, nearly 100,000 people are practicing respiratory therapy in the United States. Of those, only about 20,000 are members of the American Association for Respiratory Care. Out of all those people, less than 600 were involved with presenting research at the 47th International Respiratory Congress in 2001. Yet, every one of those 100,000 people needs to know how to read and understand scientific articles in medical journals. The same holds true for all health care workers. Even if you never conduct a study, you must be familiar with the basic concepts of research in order to practice as a professional whose understanding grows from continuing education.

The main purpose of this handbook is to help you become an educated consumer of medial research. If you want to actually perform research, the best thing you can do is find a mentor; someone who has experience conducting scientific studies and publishing the results. A mentor can help you turn the ideas in this handbook into practical realities.

THE IMPORTANCE OF RESEARCH IN HEALTH CARE

Health care professionals must acquire the knowledge and skills needed to assess the usefulness of new equipment, the effectiveness of present and proposed treatment modalities, the quality of services provided, and the adequacy of teaching materials available. *The most important of these skills is the ability to read and critically evaluate the published reports presented by other investigators.* Without this skill, no meaningful evaluation of current practices can be made and no research can be planned. The word *research* is typically used in a generic sense to mean a systematic method of inquiry.

The pursuit of scientific knowledge in any field must ultimately rest on the testing and retesting of new ideas and their practical application. Growing numbers of clinicians, educators, and administrators are conducting their own investigations and critically examining research done by others in their particular field of interest.

The experimental approach may be broken down into five phases (Table 1-1). Health care workers are usually involved with the application of research results in the clinical setting. Within the research continuum, however, an infinite number of opportunities exist to become involved in seeking the answers to questions relating to the practice of health care.

Table 1-1 The Five Phases of Research

1. *Basic Research.* Seeks new knowledge and furthers research in an area of knowledge rather than attempting to solve an immediate problem.
2. *Applied Research.* Seeks to identify relationships among facts to solve an immediate practical problem.
3. *Clinical lnvestigations.* Seek to evaluate systematically the application of research findings in the clinical setting, usually in a relatively small patient population.
4. *Clinical Trials.* Seek to determine the effectiveness and safety of clinical treatments in samples of patients drawn from larger populations.
5. *Demonstration and Education Research.* Seeks to examine the efficacy of treatments designed to promote health or prevent disease in defined populations.

The following discussion outlines several areas of health care where we may apply the principles of scientific analysis to provide a more sound basis for patient care. These include health care education, professional accountability, and administration of services.

Health Care Education

Colleges are responsible for graduating practitioners who are knowledgeable and current in the practice of their profession. Educators must stay up-to-date with new ideas and technology in medicine that affect the diagnosis and treatment of disease.

Critical Evaluation of Published Reports. Before a particular piece of equipment or treatment modality is accepted for introduction to the student, the instructor must first discern whether the claims for its use and potential benefits rest on a solid scientific foundation. Keeping abreast of new product developments requires that instructors read and critically evaluate reports and tests of function and reliability. A critical reading of scientific journals will provide the basis for their decisions concerning classroom demonstrations, guides, and the planning process. Educators may wish to conduct their own investigations as well.

The results of published reports should never be accepted uncritically. The use of intermittent mandatory ventilation (IMV), for example, was claimed to decrease the time required to wean a patient from mechanical ventilation. Yet recent studies have shown that the average length of time a patient spends on the ventilator and in the hospital actually *increased* by the use of IMV.

How much credence should we give to each of these studys' results? Is one or the other limited by its design? Does a non-uniformity of patient populations exist? Were the types of IMV systems used the same in each study? What criteria were used for judging a patient's readiness for removal from mechanical ventilation? Health care educators must ask these types of questions of all studies before passing the results on to their students; they must do more than simply take a study's conclusions at face value.

Continuing Education. In order that health care practitioners keep informed of recent developments in cardiopulmonary medicine, hospital department managers must establish and maintain continuing education programs. These inservice programs serve to explore and provide a forum for new trends, ideas, and developments that occur in the field as research is completed in varying areas of special interest. Allied health professionals are taking an increasing role in patient education as well as in clinical practice. As they are kept current on data relating to, for example, the relationship of cigarette smoking to heart disease or cancer, they can increase a patient's awareness of the appropriateness of particular treatment modalities.

The results of research on health care practices serve to reeducate practitioners and update department procedure manuals. Thus, guidelines are provided for the improvement of clinical competence. This occurs as state of the art data on equipment, care modalities, physical diagnosis, and monitoring procedures are made available and their validity tested.

Professional Accountability

Health care professionals are accountable not only to their patients, departments, and hospital administrators, but to government agencies, third-party reimbursers, and the public at large. Our nation's entire health care system is under increasing pressure to justify the cost of services it provides. Government agencies and third-party reimbursers are asking us to show that the services we provide are both necessary and beneficial.

With our country's present state of economic austerity, allocation of funds to health care agencies, such as the Federal Drug Administration (FDA) and Centers for Disease Control (CDC), has been reduced. The functioning of these agencies, as well as Medicare, Medicaid, and Blue Cross/Blue Shield, affects health care both directly and indirectly. Investment in health care for the elderly and poor by the government is under close scrutiny to make sure that funds are going to pay for justifiable services. Understandably, with an increased federal role in paying the bills, there is increased pressure to assure the quality and quantity of care and that it is cost-efficient.

The high cost of health care must be supported by scientific justification. Regulations governing medical services and reimbursement are based on the current state of knowledge. Relevant questions about a service regard its necessity for the treatment of an established medical problem and whether it is of demonstrable benefit to a patient. The task of medical officials is to assure that the appropriate regulatory body has this information at its disposal. The task of health care researchers to make certain that the information is based on scientific data.

Administration of Health Care Services

Health care department managers and hospital administrators alike look toward the results of carefully completed studies to help solve problems relating to areas of concern such as cost containment, productivity assessment, departmental organization, and employee stress management. Managers are responsible for staffing their departments with qualified personnel, providing services that are delivered in a professional and timely manner, and making certain that infection control, safety, and preventative maintenance programs are ongoing and productive. How can managers best evaluate these services and programs? Which method of providing infection control, for instance, should a manager decide on? Knowing that equipment can be a major source of nosocomial infection, a method is needed of assessing the resultant change in infection rate that a program of disinfection or sterilization will hopefully affect. The cost-effectiveness of different methods must also be taken into consideration. The same type of

questions may be asked of patient and employee safety programs, and of other organization, delivery, and evaluation of patient care.

Evaluation of the quality of departmental programs and services is a difficult challenge. Empirical observation must not be the basis for acceptance or rejection. The costs of trial and error remedies are too prohibitive for this type of decision-making.

Continuous Quality Improvement The Joint Commission on Accreditation of Health Care Organizations (JCAHO) defines quality assurance as "a manner of demonstrating consistent endeavor to deliver optimal patient care with available resources and consistent with achievable goals. The correction of deficiencies is inherent to the process." This correction process is accomplished through the careful and rigid manipulation of variables and the measurement of any effects; in other words, using the scientific method. Only in this way can the physician, patient, patient's family, hospital, and government administrator be assured the quality of cost-effective services.

Evaluating New Equipment and Methods

Validating Manufacturer's Claims. To meet the changing needs of health care, medical equipment manufacturers introduce to the market new diagnostic and support instruments. Because of the relatively short product life cycle in the market of technical equipment, new products are introduced frequently. But *new* does not necessarily mean *better*. At times, the development of new technology outpaces the need for that technology. When this happens, product marketers have not done their job in accurately assessing demand. Medical professionals must then take the lead in assuring that they are not left in the position of trying to invent ways to use new equipment. Rather, new equipment should satisfy a well-established need. Although manufacturers often engage in extensive testing and market research, the final burden of proof as to a product's ultimate function and benefit falls to the end user, us.

For example, the introduction of synchronized intermittent mandatory ventilation (SIMV) on the Bourns Bear I ventilator came about as a result of the clinical observation that some patients breathing through standard IMV systems sometimes had mandatory machine breaths delivered during exhalation or on top of their spontaneous tidal breath. This *stacking* of breaths was believed to be harmful, or at least inefficient. SIMV ensures that a mandatory machine breath is not delivered until the ventilator senses a patient's respiratory effort, thus being ready to receive a large ventilator tidal volume. Synchronizing spontaneous breathing with mandatory ventilation, it was thought, would solve the problem of stacking and encourage more efficient breathing.

In principle, SIMV makes sense. But does it make a difference in any measurable sort of way? Does it make a difference in terms of alveolar ventilation, peak airway pressure, arterial PO_2, arterial PCO_2, or patient comfort? Are the potential benefits worth the added expense of this new ventilator feature?

These types of critical questions must be asked and systematically addressed when any new piece of equipment is made available to the field. Regarding SIMV, clinical research has indicated that breath stacking is indeed not clinically significant and that hemodynamic and arterial blood gas measurements do not improve when patients are switched from IMV to synchronized IMV.

Rather than accept on faith that a new technology will do exactly what its manufacturer claims, we should validate claims and conduct comparison tests with existing equipment. We should ask questions such as: What is the chance of nosocomial infection with this equipment? Does this equipment work equally well on a patient with chronic obstructive pulmonary disease (COPD) as it does on one with a flail chest? How accurate are the pressure manometers, spirometers, and gas analyzers provided?

Empirical observations often indicate a need for a new piece of equipment or procedure. But to insure safe and effective application, its final implementation must rest on sound scientific judgment.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- Basic research
- Applied research
- JCAHO
- Quality assurance

True or False

1. The most important reason for studying research methodology is to gain the ability to read and critically evaluate studies published in medical journals.
2. The best thing you can do if you want to really learn how to do research is to find a mentor.

Multiple Choice

Which of the following are areas where we may apply the principles of scientific analysis to improve patient care:

a. Education.

b. Continuous quality improvement.

c. Evaluation of new equipment.

d. All of the above.

Chapter 2. Ethics and Research

In the health care industry today we are confronted with a multitude of laws, regulatory constraints, and standards that govern the conduct of the industry itself and the individuals who work in it. Conducting health care in this environment requires constant attention to a multitude of details. Conducting health care research demands additional attention to a special set of regulatory and ethical considerations.

Research involving human subjects, which we will refer to as *clinical* research, invokes legal, ethical, and sociologic concerns related to the safety and protection of the subject's basic human rights. Research involving animals requires responsible attention to several important concerns as well. Regardless of the type of study subjects, those engaged in medical research must be reminded that the importance of their work should never overshadow but, rather, complement society's health care goals. Complex procedures must strictly adhere to legal guidelines so that subjects are not exploited. Innovative and controversial research must be ethically conducted and honestly reported

The current and future prospects for productive and informative research in health care are as high now as they have ever been. In pursuing these prospects, the health care researcher must not only be concerned with the proper methodologies and logistics of running the actual study, but with legal and ethical issues that are no less important. Structuring research that is within the bounds of ethically and scientifically rigorous standards is an important and complex task, with a multitude of subtleties. The research investigator must achieve scientific rigor, while at the same time maintaining the highest ethical standards.

A complete discussion of all the ethical and legal implications of clinical research is beyond the scope of this text. The goal of this chapter is, first of all, to familiarize researchers with the institutional approval process they will need to navigate to begin research involving human subjects. Second, this chapter is designed to heighten the investigator's awareness with respect to several legal and ethical concerns they will undoubtedly encounter as they design and conduct their research endeavors in our modern environment. Finally, we will touch briefly on current ethical and legislative guidelines for conducting research involving animals.

INSTITUTIONAL REVIEW AND HUMAN SUBJECTS' RIGHTS

When human beings are used in scientific research, great care must be taken to insure that their rights are protected. To guarantee that protection, review boards have been established to ensure that proposed studies do not violate patient rights within a particular institution.

Functions of the Institutional Review Board

The health care researcher cannot and should not begin an investigation involving human subjects without formal approval from the hospital's Institutional Review Board (IRB). Also known as the Institutional Review Committee, Human Subjects Review Committee, Human Investigation Committee, or Research Surveillance Committee, IRB refers to any committee, board, or other group formally designated by an institution to review biomedical research involving human subjects. This committee meets at certain specified intervals to review, recommend, and approve study proposals.

The main functions of the IRB are to protect the rights, well-being, and privacy of individuals, as well as protect the interests of the hospital or center in which the research is conducted. Specific IRB procedures will vary from institution to institution. In each case, health care workers must review those guidelines applicable in their own institution.

Although IRB guidelines may vary somewhat from one institution to the next, IRBs are typically established, operated, and function in conformance with regulations set forth by the US Department of Health and Human Services (DHHS), regulations established to protect the rights of human subjects that apply to all institutions receiving federal funds. The DHHS issued regulations in 1981 that must be followed for biomedical and behavioral human research to receive such funds.

Consideration of risks, potential benefits, and informed consent typically occupies the majority of the IRB's time. Before an IRB can approve a research protocol, the following conditions must be met.

1. The risks to the (research) subject are so outweighed by the sum of the benefits to the subject and the importance of the knowledge to be gained as to warrant a decision to allow the subject to accept these risks.
2. Legally effective informed consent will be obtained by adequate and appropriate methods.
3. The rights and welfare of any such subjects will be adequately protected.

Review of research involving human subjects must always occur before the initiation of research and may be required at specified intervals during the lifetime of the research activity. If an application for external funding is being considered, the researcher should thoroughly review the study proposal before submission to the funding agency. The IRB frequently may ask the investigator to modify the original research plan to comply with Food and Drug Administration (FDA) and DHHS regulations as well ethical norms. However, the IRB is not a police force. There is a presumption of trust that the approved research protocol will indeed be followed consistently. Nevertheless, investigators have been known to deviate from the agreements reached with an IRB.

Composition of the Institutional Review Board

To provide input representing a wide variety of concerns, the IRB committee is typically composed of members with diverse backgrounds. An IRB characteristically includes representatives of administration, staff, and legal areas of both the institution and the community. This diversity encourages that proposed research be reviewed for acceptability, not only in terms of scientific standards, but in terms of community acceptance, relevant law, professional standards, and institutional regulations as well.

As well as a diverse background, committee members exhibit a high standard of personal and professional excellence. IRB members should exhibit sufficient maturity, experience, and competence to assure that the Board will be able to discharge its responsibilities and that its determinations will be accorded respect by investigators and the community served by the institution. The quality of an IRB decision is thus a direct reflection of the degree of maturity, experience, and competence of its members

Approval of the Institutional Review Board

The investigator must formally apply for IRB approval before beginning a study. A thorough IRB application typically includes the components listed in Table 2-1. First, a formal research protocol must be established. This description of the study's intended purpose and procedures is then followed by human subjects information, which should describe sources of potential subjects and the anticipated

number required. Also included should be a description of the consent procedures, and a description of potential risks and benefits as they relate to both the subjects and to society.

An integral part of the study protocol, and a necessary component for IRB review, is the patient or subject consent form, discussed in greater detail later in this chapter. To prepare this form properly, a number of issues (Table 2-1) must be thoroughly addressed. The content of each of these areas of concern must then be prepared with the consent form for the information of the potential study subject.

TABLE 2-1. Typical components of an IRB proposal.

1. A complete description of the study's intended purpose and procedures to be followed.
2. A description of potential risks the subject may incur from participation in the study.
3. A description of potential benefits, either direct or indirect, the subject may incur from participation in the study.
4. A description of how data will be handled such that the subject's identity remains anonymous.
5. A statement that the subject may withdraw from the study at any time without a prejudicial effect on his or her continuing clinical care.
6. The name and number of the investigator, should any questions arise regarding the subject's participation in the study.
7. Copy of the complete Informed Consent form.
8. A list of available alternate procedures and therapies.
9. A statement of the subject's rights, if any, to treatment or compensation in the event of a research-related injury.

INFORMED CONSENT

Informed consent is the voluntary permission given by a person allowing himself to be included in a research study after being informed of the study's purpose, method of treatment, risks and benefits.

A key principle of ethical conduct in research is that participation in studies must be voluntary. In turn, voluntary consent is predicated on communicating all the information the potential subject needs to be self-determining. The consent form represents the culmination of much effort devoted to protect the rights of research subjects through the process of fully informing them before their involvement in clinical research.

Background

The Nuremburg Trials after World War II revealed the atrocities committed by Nazi physicians. As a result of these revelations, voluntary informed consent became a central focus of biomedical ethics. The doctrine of informed consent is designed to uphold the ethical principle of *respect for persons.* As such, this doctrine is now grounded in a body of medicolegal decisions that cite a failure to obtain adequate informed consent as either *battery* or *negligence.*

Having received critical commentary for the past 35 years, the protection of human subject's rights has received formal legislative attention within the past 20 years. In legitimizing this emphasis, the World

Medical Association adopted the Declaration of Helsinki in 1964. This declaration recommended that informed consent be obtained "if at all possible, consistent with patient psychology" for "clinical research combined with patient care." Before this, potential volunteers were protected only by the assumed responsibility of the individual investigator to explain fully the nature of the research. But abuses of this responsibility led to the development and implementation of the informed consent requirement.

Role Today

Today, informed consent is a crucial feature of virtually all clinical trials. No research involving human subjects should be initiated without the informed and voluntary consent of those subjects. Competent patients must be offered the opportunity to accept or reject a medical intervention proposed as part of their participation in a research study. Likewise, incompetent patients must be offered the same opportunity through the mediation of a legal guardian or surrogate.

For consent to be *informed,* the potential subject must be given information regarding all the possible pros and cons of the proposed medical intervention. We always move toward maximizing the patients' best interests while enhancing their participation in decision-making. As a vehicle, the consent form clearly summarizes the IRB application. The consent form must contain all the elements (Table 2-1) necessary so that a patient's rights will be protected should he or she elect to participate in the research study.

Revocation of Consent

A subject may withdraw from a research activity at any time during the course of the study, within the limits of the research. Any request for withdrawal should be honored promptly. As spelled out in the consent form, revocation of consent and participation in the research study should never result in a subject being penalized or made to forfeit benefits to which he or she is otherwise entitled**.** However, the subject's commitment to participate in a research study is seen by some to represent a moral obligation. In this context, research can be viewed as a joint venture between investigator and subject. The subject has made a promise to participate, to bear the inconvenience of testing in return for the benefit he or she hopes to derive.

Nevertheless, should a subject wish to withdraw from participation in research, the investigator must fully inform the subject of the potential dangers. For example, an asthmatic subject who abruptly withdraws from a study examining the efficacy of an investigational bronchodilator should be informed of what improvement or lack thereof he or she had shown with the use of that bronchodilator. Should the subject still choose to withdraw from the study, a smooth transition to an alternative bronchodilator must be provided. If a subject suffering from pneumonia wishes to withdraw from a study of the effect of chest physiotherapy on spirometric and plethysmographic values, that person should be informed of his or her progress since the administration of the investigational treatment regimen. Alternatives to the current mode of therapy must be described so the pros and cons of withdrawal from the study can be properly evaluated. In all instances, the implications of withdrawal from therapy must be made clear to the subject, and arrangements made for a smooth, uneventful transition to standard clinical care.

ETHICAL ISSUES

Basic Principles

Professional ethics in health care ethics is a subset of the category of medical ethics, which in turn is a division of the much broader philosophy of ethics. Although the law sets a minimum level of expected behavior, ethics generally requires more than the minimum, and often aims toward the ideal. Every clinical researcher, regardless of the study, has relevant ethical responsibilities to which he or she may be held accountable. The following discussion will address ethical decisions in the field of clinical research as they concern health care investigations.

Three fundamental ethical principles relevant to clinical research are *respect for persons, justice,* and *beneficence.* Respect for persons is interpreted to mean that those conducting clinical research will endeavor to treat potential subjects as autonomous, self-determining individuals. Furthermore, those subjects not capable of making considered judgments (incompetent), those either immature or incapacitated, are entitled to the protection they deserve. The principle of justice requires that all persons be treated fairly and equally. Finally, beneficence can best be understood as a commitment to do no harm and to maximize the potential benefits while minimizing potential harms. Incumbent in this definition is the understanding that no person will be asked to accept risks of injury in the interest of producing societal benefits.

Research studies that violate these standards have been documented and serve as a basis for the contemporary balance between human experimentation and legal regulation of medical research. In 1932, male prisoners with syphilis were recruited without consent and misinformed as to their treatment. When penicillin became available for the treatment of syphilis, these men were not informed. In another study, patients with various chronic debilitating diseases were injected with live cancer cells. Consent was said to have been negotiated, but was never documented due to the investigator's contention that informing the patients of the procedure would frighten them unnecessarily. These and other abuses have combined to tighten both legal regulations and ethical guidelines for clinical research.

Ethical concepts differ substantially from legal concepts. Ethical concepts have evolved into the various professional standards and principles that guide the practice of medicine. Professional standards do not carry the weight of law; only statutes and common law have any legal authority in this country. However, many statutes and many court decisions have been, and will continue to be, extensively based on the moral and ethical convictions of the health care professions. Health care ethics may be considered a subset of the larger fields of medical ethics. No longer is the physician the absolute ruler and his or her ancillary helpers mere followers who cannot be expected to exercise any moral judgment of their own. Furthermore, medical care is no longer delivered solely by physicians and nurses. The contemporary health care industry employs a variety of professional health care practitioners, each with a high and noble ethical code of conduct no less meaningful than the Hippocratic oath. For example, the field of respiratory care operates under an ethical code, represented by the American Association for Respiratory Care Code of Ethics (Table 2-2). The issues of health care ethics are becoming more numerous and complex with nearly every major medical advance that is implemented. A partial list of the pressing issues of the current time would include death with dignity, euthanasia, discontinuation of life support systems, organ transplantation, genetic engineering, behavior modification, use of animal experimentation, and a further subset of issues that come under the general heading of human experimentation for health care research. In addition to the basic ethical principles of respect for

persons, justice, and beneficence, what other issues can the health care researcher expect to confront? There is several discussed below that deserve consideration.

Table 2-2. Statement of Ethics and Professional Conduct

In the conduct of professional activities the Respiratory Therapist shall be bound by the following ethical and professional principles. Respiratory Therapists shall:

Demonstrate behavior that reflects integrity, supports objectivity, and fosters trust in the profession and its professionals. Actively maintain and continually improve their professional competence, and represent it accurately.

Perform only those procedures or functions in which they are individually competent and which are within the scope of accepted and responsible practice.

Respect and protect the legal and personal rights of patients they care for, including the right to informed consent and refusal of treatment.

Divulge no confidential information regarding any patient or family unless disclosure is required for responsible performance of duty, or required by law.

Provide care without discrimination on any basis, with respect for the rights and dignity of all individuals.

Promote disease prevention and wellness.

Refuse to participate in illegal or unethical acts, and refuse to conceal illegal, unethical or incompetent acts of others.

Follow sound scientific procedures and ethical principles in research.

Comply with state or federal laws which govern and relate to their practice.

Avoid any form of conduct that creates a conflict of interest, and shall follow the principles of ethical business behavior.

Promote health care delivery through improvement of the access, efficacy, and cost of patient care.

Refrain from indiscriminate and unnecessary use of resources.

Objective Patient Care

Under the auspices of a physician, the health care practitioner contractually undertakes to give a patient the best possible treatment. Indeed, at the core of modem medical ethics is the Hippocratic promise to do one's best for every patient and to do no harm. Does the very act of enrolling a patient in a randomized clinical trial violate this obligation? Consider the patient with chronic obstructive pulmonary disease who agrees to participate in a study of the effects of a new bronchodilator. Does this subject fully understand the implications of falling into the placebo group? Does a subject suffering from cystic fibrosis fully understand that randomization to the control group may mean that the frequency of chest physiotherapy will not be increased during the study period regardless of a relative deterioration in his measured work of breathing. Some critics believe that if a clinician or investigator has reason to believe that the experimental treatment is better than the control treatment, he or she must recommend the experimental option. For example, suppose a new aerosolized drug seems to be highly effective and superior to the standard treatment of patients with acute respiratory distress syndrome. A controlled clinical trial is undertaken, with 50 patients randomized to receive conventional therapy of mechanical ventilation with positive end-expiratory pressure, increased FiO_2 and supportive fluid therapy. Another 50 patients are randomized to the treatment group, and receive the new drug in addition to conventional therapy. Now suppose that 15 patients in the experimental group die, as opposed to 30 patients in the control group. Is the clinical investigator guilty of unethical behavior? Is he or she guilty of a crime, a sin of omission?

Unfortunately, there are no clear-cut answers. As is made clear in the Nuremberg codes, the degree of risk to be taken should never exceed that determined by the humanitarian importance of the problem to be solved by the experiment. In other words, there should always be a favorable balance between harm and benefit. The Declaration of Helsinki further reinforces this principle in stating that "Biomedical Research involving human subjects cannot legitimately be carried out unless the importance of the objective is in proportion to the inherent risk to the subject." The fundamental ethical principle is that of beneficence. Furthermore, justice and respect for persons are served when a study's potential harms and benefits are clearly and properly presented to the subject for his or her informed consent

Reporting Research Results

Scientific investigations are based to a very high degree on trust. We trust that each investigator will conduct his or her research in accordance with the protocol approved by the appropriate IRB. And we trust that all research findings will be reported accurately and without intentional bias. Abandoning trust would lead to overwhelming suspicion and make scientific investigation impossible. Without trust in the honesty and integrity of published findings, how would progress in science and medicine be possible?

Fortunately, fraud in science is rare, due to the skepticism of the scientific community. No experiment is accepted until it has been independently repeated. Research results, no matter how sensational, are quickly forgotten if they cannot be obtained from other investigators duplicating the study methodology.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- IRB
- Informed consent

True or False

1. The IRB is intended to protect the rights of patients involved in research studies.
2. The IRB is composed of the people who designed the research study.

Multiple Choice

1. Typical components of an IRB proposal include:
    a. Description of study purpose
    b. Potential risks and benefits
    c. Informed consent form
    d. Description of investigator's previous experience
    e. All of the above.
    f. Only a, b, and c.
2. Three fundamental ethical principles relevant to clinical research are:
    a. Respect for persons.
    b. Cost containment.
    c. Justice.
    d. Beneficence.
    e. a, b, d
    f. a, c, d
3. At the core of modern medical ethics is the Hippocratic Oath, which obligates caregivers to:
    a. Treat everyone fairly.
    b. To do no harm.
    c. To give only treatment proven by scientific methods.
    d. To obtain informed consent before entering a person in a study.

Chapter 3. Outcomes Research

As in other areas of medicine, outcomes research is starting to make its mark in defining optimal health care practices. With the need for cost containment, outcomes research becomes a double-edged sword used both to cut nonessential practices and to protect those that maintain quality of care. The profession of health care has a long history of research and a commitment to basing practice on science. However, much of the published research is still focused on devices and procedures rather than the broader issues of patient outcomes and economic effects. We need to evolve our paradigms to accommodate the larger vision of disease management, which encompasses the arenas of outcomes research and evidence-based medical practice.

In this chapter, I will give a brief history of the outcomes research movement to provide some sense of context. Then, I will try to demystify the language of outcomes research and review some of its themes and methods. Finally, I will present specific examples of outcomes research found in the pages of Respiratory Care journal. Hopefully, these examples will illustrate some of the methods of outcomes research and stimulate future studies.

A BRIEF HISTORY

Florence Nightengale may have been the first outcomes researcher in medicine. She had a flair for collecting, analyzing and presenting data. She even invented the polar-area chart, where the statistic being represented is proportional to the area of a wedge in a circular diagram. Yet, she had as much trouble finding appropriate data as we do today. And like modern times, there was much opposition to the reforms proposed by Nightengale. Nevertheless, her most effective weapon was the presentation of solid, relevant data. For example, she showed "...that 'those who fell before Sebastopol by disease were above seven times the number who fell by the enemy.' The opposition could not respond to her statistics and publication of the statistics led to public outcry."

The modern outcomes movement in the United States had its beginnings in the early 1980s. The increasing focus on cost containment led to interest in identifying and eliminating unnecessary procedures. Perhaps more intriguing was the recognition that there were substantial variations in medical practice, apparently based on geography or race. Indeed, some researchers claimed that "geography was destiny" because medical practices as commonplace as hysterectomy and hernia repair were performed much more frequently in some areas than in others, with no differences in the underlying rates of disease.

Given that there are variations in practice and differences in outcomes, we may logically assume that some practices produce better outcomes than others. So the stage was set to improve efficiency and quality if only the right data were available. But where to look? The Office of Technology Assessment estimated that only 10% to 20% of interventions by physicians have been clearly shown in randomized clinical trials to be of value. This is not surprising, given that clinical trials can cost millions of dollars and last years. Some suggested that data collected for administrative or billing purposes (e.g., Medicare and Medicaid tapes collected by the Health Care Financing Administration) might contain valuable outcome data such as mortality, length of hospital stay, resource use, and costs. On the one hand, such data can be quickly analyzed, without requiring patient consent or interfering with medical care. On the other hand, critics argued that this type of research is limited by the quality and completeness of the data.

New data must be collected in a systematic fashion with a specific focus on outcomes. In 1989 Congress created the Agency for Health Care Policy and Research (AHCPR). It consisted of 11 major components including the Center for Outcomes and Effectiveness Research, the Center for Cost and Financing Studies, and the Center for Quality Measurement & Improvement. The initial focus of the AHCPR was to create Patient Outcomes Research Teams (PORTs; 5-year studies of specifically identified diseases with highly focused methods), the Pharmaceutical Outcomes Research Program, and the Minority Health Research Centers. In time, the AHCPR changed its name and its focus. Today, at the Agency for Healthcare Research and Quality, the focus is on Translating Practice Into Research, creating Excellence Centers for Eliminating Disparities (based on race and ethnicity) and supporting the Centers for Education and Research on Therapeutics. According to the AHRQ, the purpose of outcomes research is to answer four basic questions:

- What works?
- What doesn't?
- When in the course of an illness (does it work or not)?
- At what cost.

These questions suggest the scope and focus of modern outcomes research.

UNDERSTANDING THE JARGON

Like any new discipline, the field of outcomes research suffers from a lack of consistent definitions and a unifying conceptual framework. Many seemingly unrelated terms are encountered in the literature such as efficacy, effectiveness, quality of life, patient centered care, evidence based medicine, etc. All these terms signify a paradigm shift in which the emphasis is on populations rather than individuals; on practice guidelines rather than anecdotal justifications for treatment; and on capitation rather than fee-for-service payments. I have found it helpful to view this new paradigm in terms of the general concept of "disease management" within which the specific activities of outcomes research and evidence-based medicine interact in a process of continuous quality improvement.

Disease management (also called outcomes management) can be defined as the systematic, population based approach to identify patients at risk, intervene with specific programs, and measure outcomes. The basic premise of disease management is that an optimal strategy exists for reduced cost and better outcomes. Disease management emphasizes identifying populations of interest, creating comprehensive interventions, explicitly defining and measuring outcomes, and providing a strategy for continuous quality improvement.

Continuous quality improvement (CQI) is a cycle of activities focused on identifying problems or opportunities, creating and implementing plans, and using outcomes analysis to redefine problems and opportunities. CQI was started decades ago by pioneers such as Shewert, Deming, and Juran and is currently embraced by the Joint Commission on Accreditation of Healthcare Organizations. The "plan, do, check, act" cycle endorsed by JCAHO can be viewed as simply creating plans and implementing them. The plan leads to implementation through the creation of specific goals. Implementation leads to more plans through the analysis of results (Figure 3-1).

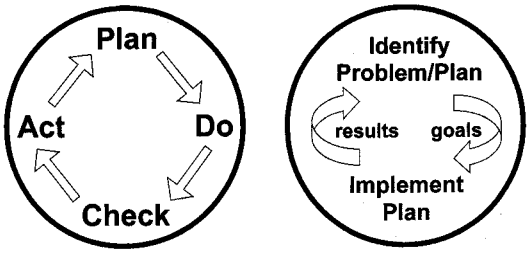


Figure 3-1. Continuous quality improvement expressed in the traditional format of a cycle of "plan, do, check, act" and an equivalent cycle showing the interaction of plans and implementations through goals and measured results.

Outcomes research can be defined as the scientific study of the results of diverse therapies used for particular diseases, conditions, or illnesses. The specific goals of this type of research are to create treatment guidelines, document treatment effectiveness, and to study the effect of reimbursement policies on outcomes.

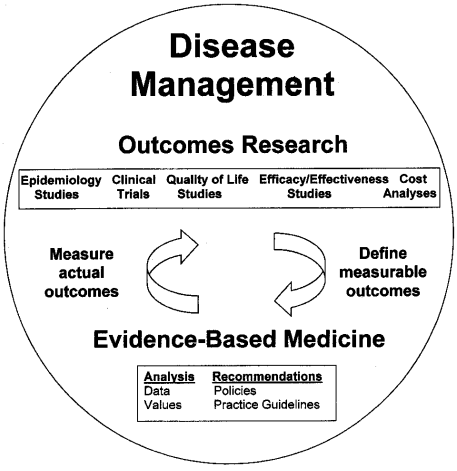
Evidence-based medicine is an approach to practice and teaching that integrates pathophysiological rationale, caregiver experience, and patient preferences with valid and current clinical research evidence. To implement evidence-based medicine, the practitioner must be able to define the patient problem, search and critically appraise data from the literature and then decide whether, and how, to use this information in practice.

If we view disease management as a universe of activities, then outcomes research (e.g., epidemiological studies, clinical trials, quality of life surveys, efficacy and effectiveness studies, and cost analyses) and evidence-based medicine (e.g., creation and use of practice guidelines and care paths) can be seen as subset activities linked by the general structure of continuous quality improvement (Figure 3-2).

OUTCOMES RESEARCH: FOCUS AND METHODS

Outcomes research can be distinguished from traditional clinical research more by its focus than on the methods it employs. This difference in focus is highlighted in Table 3-1. Appropriate outcomes can be roughly grouped into three categories; clinical, economic, and humanistic (Table 3-2).

Figure 3-2. Disease management expressed as a continuous quality improvement cycle showing the interaction of plans (created by outcomes research) and implementations (evidence-based medicine tools) through goals (desired outcomes) and measured results (actual outcomes).



Outcomes research uses a variety of techniques (Table 3-3). *Qualitative research* often produces large amounts of textual data in the form of transcripts and observational field notes. Rather than trying to identify a statistically representative set of observations, qualitative research uses analytical categories to describe and explain social phenomena. Qualitative research generates hypotheses (although not necessarily hypothesis tests), and attempts to identify the relevance of findings to specific groups of people.

Table 3-1. Differences between traditional clinical research and outcomes research.

Traditional Clinical Research	Outcomes Research
Disease-centered	Patient and community centered
Drugs and devices	Processes and delivery of care
Experimental	Observational
Methods from "hard sciences" (physics, biochemistry)	Methods from "social sciences" (economics, epidemiology, etc)
Efficacy	Effectiveness
Mechanisms of disease	Consequences of disease on patients
Effects of biochemical and physiologic factors	Effects of socioeconomic factors

Quantitative research uses both experimental and non-experimental designs. The classic experimental design of the randomized controlled trial can be applied to outcomes research if it is set up to evaluate effectiveness (as opposed to efficacy, see definitions below). Non-experimental designs can focus either on data synthesis or observational study designs.

In keeping with the population-based theme of outcomes research, methods are needed to synthesize data from numerous studies, as opposed to interpreting the results of a single study. One such method is called *meta-analysis*. The National Library of Medicine defines meta-analysis as follows:

A quantitative method of combining the results of independent studies (usually drawn from the published literature) and synthesizing summaries and conclusions which may be used to evaluate therapeutic effectiveness, plan new studies, etc, with application chiefly in the areas of research and medicine. The method consists of four steps:

1. A thorough literature review,
2. Calculation of an effect size for each study,
3. Determination of a composite effect size from the weighted combination of individual effect sizes
4. Calculation of a fail-safe number (number of unpublished results) to assess the certainty of the composite size

Table 3-2. Various types of outcome measures used in outcomes reserach**.**

Category	Type	Example
Clinical	Clinical events	Myocardial infarct
	Physiologic measures	Pulmonary function indices
	Mortality	Asthma deaths
Economic	Direct medical costs	Hospital and outpatient visits
	Indirect costs	Work loss, restricted activity days
Humanistic	Symptoms	Dyspnea scores
	Quality of life	SF-36 Questionnaire, St. Georges Respiratory Questionnaire
	Functional status	Activities of daily living
	Patient satisfaction	Cleveland Health Quality Coalition

Decision analysis is used to determine optimal strategies when there are several alternative actions and an uncertain or risk-filled pattern of future events. This technique is a derivative of operations research and game theory. It involves identifying all available choices and the potential outcomes of each. Usually a model is created in the form of a decision tree. The tree is used to represent the strategies available to the clinician and the likelihood that each outcome will occur if a particular strategy is chosen. The relative value of each outcome can also be described.

There are several basic types of economic evaluations that are applied to health issues. *Cost identification* is simply the description of the costs of providing the intervention. It is the first step in all the other types of analyses, but is often the only one reported in a study. *Cost of illness* analysis estimates the total cost of a disease or disability to society (e.g., heart disease costs the United States $128 billion per year). *Cost minimization* is applied when two or more interventions are being compared on the same outcomes and the outcomes seem to yield similar effectiveness. Then the question is simply, which is least expensive. An example would be the question of whether to repair or replace a mechanical ventilator. When the same outcomes are measured but the effectiveness differs, then they are compared on the basis of cost per outcome (e.g., dollars per life saved or dollars per additional year of life) using *cost effectiveness* analysis. If both outcomes and effectiveness differ, then a *cost-benefit analysis* first attempts to express both outcomes and benefits in terms of dollars. Then the interventions are evaluated in terms of the overall economic tradeoffs among them. In this way the cost of, for example, a smoking prevention program can be compared to that of lung reduction surgery and both can be compared to other programs such as highway development or job training. *Cost utility* analysis is similar to cost

effectiveness except that the effectiveness is expressed as a "utility" which is the product of a clinical outcome, such as years of life saved, and a subjective weighing of the quality of life to be had during those years. Utility is often expressed as quality-adjusted life years (QALYs). For example, quality of life is often measured on a linear scale where 0 indicates death (or indifference to death) and 1.0 represents perfect health. Suppose a patient is discharged to a chronic ventilator weaning facility for 6 months and dies on the ventilator. If the patient rates the utility of life on the ventilator as 0.2, the patient has experienced 0.5 x 0.2 = 0.1 QALYs. If the assumptions are correct, this means that 6 months on a ventilator in a weaning facility is approximately equal in value to the patient as 1 month (0.1 year) in perfect health. Economic analyses can seem overly complicated. For a very readable introduction written in the style of a conversation between two doctors, see the article by Eddy.

Table 3-3. Methods used in outcomes research.

- **Qualitative methods** (formal hypothesis testing not necessarily required)
  - Generate hypotheses
  - Describe complex phenomena
  - Identify relevance of findings to specific patient groups/
- **Quantitative Methods**
  - *Experimental*
    - Randomized controlled trials
  - *Non-experimental*
    - Data synthesis
      - Meta-analysis
      - Decision analysis
      - Economic analysis
  - *Observational studies*
    - Cohort
    - Case-control
    - Survey

Quality of life (QOL) measures have been important in research since the 1970s. Uses of QOL data include distinguishing patients or groups, evaluating therapeutic interventions, and predicting patient outcomes. However, there are many QOL instruments and much theory but no unified measurement approach. And there is little agreement on definitions and interpretations. Some authors argue that because QOL is a uniquely personal perspective, patient-specific measures should be used.

Another issue that seems confusing is the difference between efficacy studies and effectiveness studies. An example of the type of question answered by an *efficacy* study is as follows: "Does the intervention

work in a tertiary care setting with carefully selected patients under tightly controlled conditions?" This type of study generally requires a priori hypotheses, randomization of subjects to predefined treatments, homogeneous patient populations at high risk for the outcome, experienced investigators following a specific protocol, a comparative intervention (e.g., a placebo) and intensive follow-up. Conclusions from this type of study prompt relatively high levels of confidence. However, because the design is so restrictive, the results may not be generalizable to a broad range of patients in usual practice settings. Thus, efficacy studies may not be appropriate for cost-effectiveness analyses.

In contrast, *effectiveness* studies are designed to answer questions such as: "Does the intervention work in clinical practice settings with unselected patients, typical care providers and usual procedures." Many effectiveness studies have been conducted as observational (often retrospective) studies where observed groups were not randomly assigned and neither patients nor providers knew they were being studied. The weakness of this study design is that selection bias may be a problem (i.e., the groups may not have the same prevalence of confounding variables) so adjustment for factors such as severity of illness and case mix becomes important. Prospective effectiveness trials have been reported. They differ from typical clinical trials in that they enroll heterogeneous participants, impose few protocol-driven interventions, and report outcome measures relevant to the delivery system.

THE OUTCOME OF OUTCOMES RESEARCH

The Lewin Group has created a report of outcomes and effectiveness research (OER) at the Agency for Health Care Policy and Research that describes the accomplishments and lessons of the past decade. The report describes a conceptual framework for understanding and communicating the impact of OER on health care practice. Four levels of impact are defined:

1. Findings that contribute to but do not alone reflect a direct change in policy or practice, such as new analytic methods or outcome instruments.
2. Research that prompts the creation of a new policy or program, such as an AARC clinical practice guideline.
3. A change in what clinicians or patients do.
4. Actual changes in health outcomes.

A survey was mailed to all principal investigators (PIs) funded by AHCPR's Center for Outcomes and Effectiveness Research between 1989 and 1997. The results suggest that PIs have been most successful in (a) providing detailed descriptions of what actually occurs in health care, (b) developing tools for measuring costs of care and patient reported outcomes, and (c) identifying topics for future research. Few PIs reported findings that provide definitive information about the relative superiority of one treatment strategy over another. Furthermore, there were few examples of findings that have been incorporated into policy (level 2 impacts) or clinical decisions (level 3), or interventions that have measurably improved quality or decreased costs of care (level 4). The report concludes that "*One of the main challenges for the next generation of outcomes studies is to move from description and methods development to problem solving and quality improvement*."

I should point out that not everyone believes that outcomes research is good. Some authors voice both practical and philosophical arguments against the outcomes movement. They claim the outcomes movement exaggerates its usefulness by understating several difficulties. For example, how much time and money will be required to determine the effectiveness of many commonly used (and continuously evolving) medical procedures? How will physicians use outcomes data when making multiple

consecutive decisions in the rush of daily patient care? And how will compliance with practice guidelines be enforced? Some data suggest that clinical practice guidelines have been remarkably unsuccessful in influencing physician behavior. Reasons for this include the fact that some guidelines are not written for practicing physicians, the issue of physician disagreement with or distrust of guidelines written by so-called national experts, and physicians choosing to ignore guidelines because of non-clinical factors such as financial incentives or fear of malpractice litigation. This last issue is echoed by the opinion that many physicians are opposed to the kind of micromanagement and attendant loss of clinical autonomy-envisioned by the participants in the outcomes movement. Some proclaim that uncertainty and subjectivity are at the heart of the clinical encounter and this will always be the case. Also, by criticizing the uncertainty of physicians, the outcomes movement may set the unrealistic goal of creating important certainties for practitioners and thereby misrepresents the terms of the clinical encounter and inadvertently undermines confidence in the physician's ability to act wisely in the face of inevitable uncertainty.

EXAMPLES FROM RESPIRATORY CARE

Outcomes research, along with its methodologies and core curriculum, can be viewed as an important discipline for the field of respiratory care. Specific areas where outcomes research techniques could be employed include:

1. Determining the effectiveness of CQI initiatives.
2. Comparing variations in respiratory care practices in order to identify optimum strategies.
3. Developing and assessing innovations.
4. Evaluating resource utilization in areas employing respiratory care professionals compared to similar settings without them.

Indeed, the profession's scientific journal, Respiratory Care, has published a substantial amount of outcomes research in the last few years. A quick survey of articles in the Original Contribution category of Respiratory Care from 1997 through 2000 showed about 28% of articles could be classified as outcomes research. While, the majority of articles are still focused on devices and procedures, a number of those focused on problem solving and quality improvement may serve as examples. What follows is a brief description of the methodology used in a few of these studies:

Stoller JK, Orens D, Ahmad M. Changing patterns of respiratory care service use in the era of respiratory care protocols: An observational study. Respir Care 1998;43(8):637-642.

This was an observational study that qualifies as outcomes research because it was an evaluation of clinical outcomes in a "real world" setting during variations in respiratory therapy practices. The authors hypothesized that the use of a respiratory care consult service would decrease over-ordering of respiratory care services and decrease the volume of respiratory care services delivered. Data were obtained from departmental management information system (Clinivision, Puritan-Bennett) and from the hospital's cost management software (Transitions Systems). They compared baseline data from 1991, prior to establishment of a respiratory care consult service in 1992 to clinical data from 1996. Results were reported using descriptive statistics (averages, percentages, and trend graphs) of numbers of therapies, numbers of patients treated, and costs of therapies.

Adams AB, Shapiro R, Marini JJ. Changing prevalence of chronically ventilator-assisted individuals in Minnesota: Increases, characteristics, and the use of noninvasive ventilation. Respir Care 1998;43(8)643-649.

This is an example of an epidemiology study. Such studies describe the distribution and size (prevalence and incidence) of disease problems in human populations. The authors developed a study question, "Did cost constraints and changes in care settings and techniques affect the number of ventilator-assisted individuals (VAI), their sites of care, or methods used for ventilatory assistance." They defined VAIs and specified inclusion/exclusion criteria. Data were generated from surveys sent to all sites providing VAI care. Results were reported using descriptive statistics (averages, medians, percentages, and bar graphs) numbers of patients treated and diagnostic categories.

Myers TR, Chatburn RL, Kercsmar CM. A pediatric asthma unit staffed by respiratory therapists demonstrate positive clinical and financial outcomes. Respir Care 1997;43(1):22-29.

Here is an example of a controlled clinical trial designed as an effectiveness study (i.e., a heterogenous patient population treated by usual caregivers in a standard acute care environment). The authors tested the hypothesis that using respiratory therapists in a disease management model using a dedicated asthma unit and a standardized treatment protocol would improve efficiency of care compared to the historic method of random placement of patients with care dictated by individual physician preference. An algorithmic treatment protocol was defined. Patients treated by protocol were compared to historic controls. Data were obtained from patient charts and the hospital information system. Outcomes were stratified by an asthma severity index. Results were reported using inferential statistics to compare hospital length of stay, cost/case and care path variances. Nonparametric tests were used to assure the comparability of the two treatment groups on confounding factors such as age, race, and distribution of disease severity.

Parker, Walker. Effects of a pulmonary rehabilitation program on physiologic measures, quality of life, and resource utilization in an HMO setting. Respir Care 1998;43(3):177-182.

This study provides a good example of how to assess quality of life (QOL) issues. The researchers created a priori hypotheses and described the study population based on diagnosis and physiologic measures. They defined the intervention as rehabilitation classes at a specific frequency and duration along with an exercise program. Their QOL survey was abstracted from other published, validated QOL instruments. They used inferential statistics to compare charges and QOL scores. Results were reported using graphs and mean values.

Smith KJ, Pesce RR. Pulmonary artery catheterization in exacerbations of COPD requiring mechanical ventilation: A cost-effectiveness analysis.

Despite its title, this study is an example of cost-utility analysis, as I have defined previously, because the results are expressed in terms of patient utility (using a QOL score on a scale of 0 = death to 1.0 = perfect health) and quality-adjusted life-years. This article provides an excellent description of a complex topic, showing how a decision tree model is constructed (Figure 3-3), how probabilities of different outcomes are estimated, how costs are attributed and how utility is calculated. In addition, it provides an example of how sensitivity analysis is used to evaluate the effects of varying baseline values (i.e., assumptions) within the model.

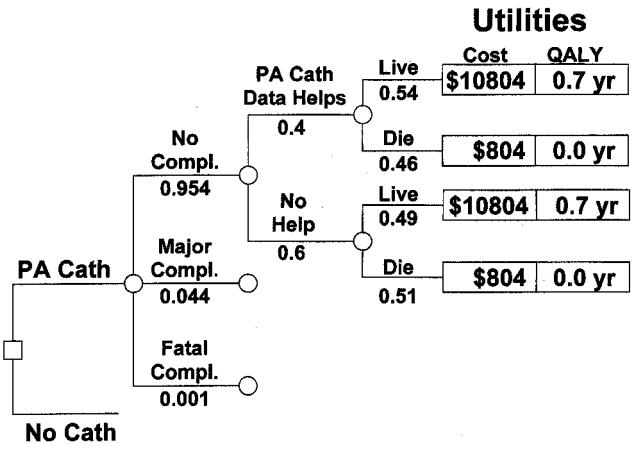


Figure 3-3. A portion of a decision tree used in a cost-study analysis. The model includes baseline values for probabilities, costs, and quality-adjusted life years (QALY). Probabilities are expressed as decimal numbers below the tree branch labels. The square node represents the decision whether to perform pulmonary artery catheterization (PA Cath) or not (no Cath). The circular nodes are the possible outcomes. QALY = life expectancy x quality of life utility value.

BENCHMARKING

Most of the medical procedures we practice each day have never been and never will be supported by formal scientific research. There simply is not enough time or money to do so. However, we can still logically justify what we do. The next best thing to scientific research is *benchmarking*. A benchmark is literally a standard or point of reference in measuring quality. As it relates to industry or health care, benchmarking is the process of comparing your performance with your peers to see who is the most successful. Benchmarking is often defined as a continuous process of measuring products, services, and practices against one's toughest competitors or renowned industry leaders, and then learning from them.

Three types of benchmarking are generally recognized: collaborative, functional, and internal. Collaborative benchmarking enables an organization to learn from the best practices within a voluntary network of health care providers. Collaborative benchmarking is often managed by a professional organization such as the University Hospitals Consortium.

Functional benchmarking compares a work function with the functional leader even when the leader is in a different industry. However, clinical functions, by their technical nature, restrict the search for benchmarking partners to health care organizations.

Internal benchmarking involves the identification of best practices within one's own organization. Internal benchmarking is both an effort to improve performance and a low risk way to share performance data. By publishing your performance data in medical journals, others can learn what a high-performing organization is doing to achieve results.

Benchmarking depends on the disciplined collection and use of objective information. The paradigm is simple, and entails:

- Identifying critical success factors and determining key indicators;
- Collecting information relevant to the key indicators;
- Searching to identify extraordinary performers, as defined by the data;
- Identifying the factors that drive superior performance;
- Adopting or adapting those factors that fit into your processes.

Benchmarking indicators are of three types:

Ratio Indicators: Indicators that establish a relationship between two measures (e.g., worked hours/unit of service). Ratio indicators are generally indicative of productivity or of a volume measurement. They provide a comparative performance point to other departments or hospitals, but do not reveal information about the practices that drive the performance.

Process Indicators: Indicators that measure a process with a beginning point and an ending point (for example, blood gas measurement turn-around-time). Process indicators lead to investigations of the practice that drives the performance.

Outcome Indicators: Indicators that measure clinical outcomes (for example, patient returns to the emergency department within 24 hours). Outcome indicators lead to an understanding of the practices that provide the best possible clinical outcomes.

Once the key indicators have been identified, useful information (data) about existing processes are collected. Most quality improvement tools depend on accurate data. The methods of data collection (except perhaps financial data) are not much different from the methods of formal research. Keep in mind that when one attempts to compare data from one department to another, the comparison is impossible unless the measures are defined in such a way that you are comparing "apples to apples".

Once data are gathered, they are analyzed using the same procedures as those used in formal research projects. These procedures include both descriptive and inferential statistics and graphical illustrations. In benchmarking jargon, this phase is sometimes called "gap analysis" because you are trying to identify any gaps or differences among benchmarking participants.

Once the gap analysis is complete and the results are known, individuals typically respond in one of three ways: denial, rationalizing, or learning.

Seldom will the results of a benchmarking project proclaim any department "best of class" across the board. More often, the news is less than uplifting, and perhaps, even discouraging. The natural response from a manager is "These data can't be correct." Unfortunately, they probably are. Facing reality is often the most difficult part of benchmarking.

The second response is rationalization. In the attempt to explain away the gaps identified in the data analysis, managers usually try to find errors in the data or methods used to collect the data. If an error can be uncovered, then they think business can continue as usual. The cry is often "We're unique!" and the implication is that just because a methodology worked in Hospital A does not mean that it will work for us because we are different.

Learning is the third response. Learning comes from accepting reality and taking actions to change it. Corrective action begins with accepting that the benchmarking data are probably correct, asking the right questions, and realizing that lessons can be learned.

The overriding objective of benchmarking is to identify and learn about best practices. But unless we implement the best practices, we have engaged in nothing more than an intellectual exercise with little value.

SUMMARY

Outcomes research seeks to understand the end results of particular health care interventions. End results include effects that people experience and care about, such as change in ability to function. In particular,

for individuals with chronic conditions (where cure is not always possible) outcome results include quality of life as well as mortality. By linking the care people get to the outcomes they experience, outcomes research has become the key to developing better ways to monitor and improve the quality of care.

The methods of outcomes research vary significantly from those of traditional clinical research. Health care workers need to be familiar with these methods be educated consumers of (and to participate in) future studies.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- Disease management
- Continuous quality improvement
- Outcomes research
- Evidence-based medicine
- Benchmarking

True or False

1. Qualitative research uses classical experimental designs whereas quantitative research relies on textual data in the form of observational field notes.
2. Outcomes research is centered on patients and communities while traditional clinical research is disease-centered.
3. Two types of *clinical* measures used in outcomes research are patient symptoms and quality of life.
4. One of the main challenges for outcomes studies is to move from description and methods development to problem solving and quality improvement.
5. *Efficacy* studies attempt to answer the question "Does the intervention work in a tertiary care setting under controlled conditions" while *effectiveness* studies attempt to answer the question " Does the intervention work in clinical practice settings."

Multiple Choice

1. An economic evaluation that is applied when two or more interventions are compared on the same outcomes and the outcomes have similar effectiveness is:
    a. Cost identification.
    b. Cost minimization.
    c. Cost effectiveness.
    d. Cost utility.

2. An economic analysis used when the same outcomes are measured but effectiveness differs is:
    a. Cost identification.
    b. Cost minimization.
    c. Cost effectiveness.
    d. Cost utility.
3. The main value of benchmarking is that;
    a. It is a practical alternative when there is not enough time or money for a scientific study.
    b. It is better than continuous quality improvement.
    c. No patient data are needed.
    d. Many hospitals can collaborate.
4. A benchmarking indicator that establishes a relationship between two measures such as worked hours per unit of service is called a:
    a. Process indicator
    b. Ratio indicator
    c. Outcome indicator
5. Common responses of managers confronted with benchmarking results include all but:
    a. Arguing that the data are incorrect.
    b. Attempting to explain away results by asserting that their situation is unique.
    c. Learning from the experience of others.
    d. Insisting on performing a gap analysis.

SECTION II PLANNING THE STUDY

Chapter 4. The Scientific Method

Research attempts to find answers using the *scientific method*. Science is simply organized curiosity. The scientific method is the organizational structure by which we formulate questions and answers during experiments. The key purpose of this organizational structure is to allow experiments to be repeated and thus validated by us or other researchers. In this way, we develop confidence in our findings. Contrary to popular belief, science does not attempt to *prove* anything. You can never prove the truth of an assumption simply because you can never test all the factors that could possibly affect it. Scientific theories are never "true", they are simply useful to various degrees and their life spans are inversely proportional to the amount of research done on them.

THE SCIENTIFIC METHOD

The scientific method is usually thought of as a series of steps that lead from question to answer (and then usually to more questions).

1. Formulate a Problem Statement

Research projects usually start out as a vague perception of some problem, either real or imagined. The first step is to refine this vague notion into a concise statement, usually only one or two sentences in length. Think in terms of (a) what you see happening and (b) why it is important. For example, if you find a coin lying on the ground your problem statement might be " I need to identify this coin so I can decide whether or not to spend it."

2. Generate a Hypothesis

The hypothesis is a short statement that describes your belief or supposition about a specific aspect of the research problem. The hypothesis is what you test with an experiment. Nobody knows where hypotheses come from; forming one is a creative act. All you can do is prepare yourself by thoroughly studying all aspects of the problem so your mind becomes a fertile ground for hypotheses to grow. Continuing with our example, we might hypothesize that "The coin is a penny." (see Figure 4-1).

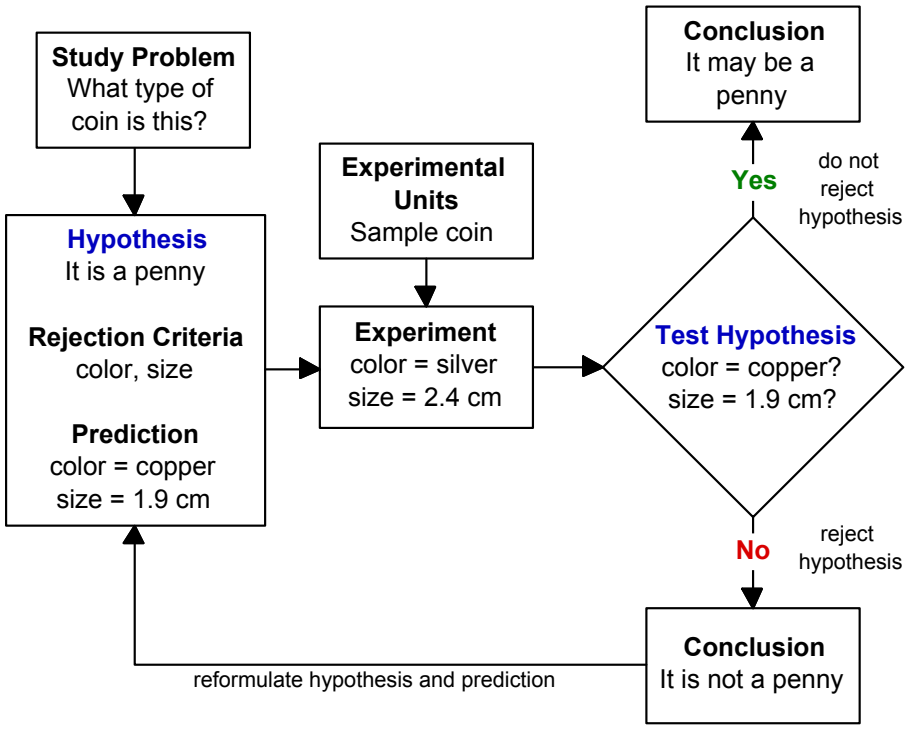


Figure 4-1. Algorithm illustrating the scientific method.

3. Define Rejection Criteria

The purpose of the experiment is to provide data. We will use the data to either reject the hypothesis as false or else accept it for the time being as a useful assumption. The fact that we can never prove the truth of a hypothesis leads us to focus on trying to prove it false. We prove a hypothesis is false by comparing the experimental data to a set of criteria we have established before the experiment began. If the experimental data do not meet the criteria, we reject the hypothesis (hence the name "rejection criteria"). In order to define the rejection criteria, we need to specify what we can measure during the experiment. For example, we could measure the coin's diameter and note its color.

4. Make a Prediction

Next, we make a prediction based on our hypothesis that specifies the rejection values. For example, we can say "If the coin is a penny, it will have a diameter of 1.9 centimeters and a copper color." The rejection criteria are thus: diameter = 1.9 centimeters and color = copper.

5. Perform the Experiment

The rejection criteria determine the measurements that are required in the experiment. Many factors are involved with designing experiments, some of which we will discuss in the next section. Much of experimental design is based on statistical theory, which is beyond the scope of this handbook. However, the basic idea is to determine (a) what variables to measure, (b) how the measurements should be made, and (c) what experimental units (subjects) will be used for making measurements. In our simple example we have only one experimental unit (the coin) and we need only a ruler and our eyes for making the measurements.

6. Test the Hypothesis

It is the hypothesis, not the experimental subject, which is being tested (despite the fact that we say things like "The patient was tested for cystic fibrosis."). The hypothesis is tested by comparing the experimental data to the rejection criteria. If the data contradict the prediction we made, then the hypothesis is rejected. If not, the hypothesis is accepted as possibly true until further data can be obtained. For example, suppose that the diameter of the coin is 2.1 centimeters and it is silver. Obviously, we would reject the hypothesis that it was a penny. We would then create a new hypothesis (perhaps that the coin was a dime) and a new prediction (based on the diameter and color of a dime).

But suppose the diameter is indeed 1.9 centimeters and the color is copper. Does that mean it is definitely a penny? What if there is a foreign coin that just happens to have those characteristics? So, we simply acknowledge that we may be wrong but until we have further information, we will suppose the coin is a penny. This example shows that we can do everything right in terms of following the scientific method and still end up with a wrong conclusion. It also shows the critical nature of selecting the right rejection criteria and making accurate measurements. It also show how science usually produces more questions than it answers.

STEPS IN CONDUCTING SCIENTIFIC RESEARCH

We will now expand on the basic scientific method to give an overview of the entire process of conducting a research project. Each step in the process will be explained further in later sections of this handbook.

Develop the Study Idea

The first step is to develop your ideas about the study problem and the specific hypotheses. Ideas come from everyday work experiences, talking with colleagues, and reading professional journals. You must also consider the feasibility of actually conducting the experiment. A great project that you do not have the resources to finish is a waste of time.

Search the Literature

An important step in the research process is a thorough search of the literature. A literature search helps you to find what has is already known about your subject and provides ideas for methods you might use for experiments.

Consult an Expert

Before you begin writing the plan for your project, discuss your ideas with someone who has experience with research and statistics. Advice at this point can help you refine the study question, identify appropriate experimental methods, and develop an implementation plan.

Design the Experiment

Three basic study designs are commonly used in health care; (1) the case study, (2) the device or method evaluation, and (3) the original clinical study. The case study is a description of a particular patient care episode that has exceptional teaching value. There is usually no need for statistical analysis so the case study may be a good choice for the novice researcher.

A device or method evaluation has at least some descriptive statistics and may even involve hypothesis testing to determine the efficacy of a treatment or compare the performance of a new device with older

devices. While this design is a little more complicated than the case study, it is very popular among new researchers because it usually does not involve the hassle of working with patients.

A clinical study is the most advanced design and usually involves sophisticated statistical procedures, medical equipment, patient care, permission to perform studies, and a variety of other logistical complications. Clinical practice is based on this type of research. However, you should not attempt this type of research until you have some experience and a good mentor.

Write the Protocol

A brief but detailed research protocol serves as a set of instructions for investigators. It also serves to communicate your plans to others, like those from whom you must obtain cooperation or permission to conduct the study.

Obtain Permission

Before conducting a study, you need permission from your immediate supervisor and from any others who will be affected (physicians, nurses, staff, lab personnel, etc). If the study involves the potential for risk to study subjects, the research protocol will have to be approved by the hospital's *institutional review board (IRB),* which is sometimes called the committee for protection of human subjects. If your study involves the medical treatment of patients (or even animals) you will probably have to get a physician to act as principal investigator in order to obtain permission from the IRB. In addition, any such study requires written consent from the study subjects or their guardians. The decision to participate in a study must be voluntary and the subject must be allowed to withdraw ant any time.

Collect the Data

The best-laid plans often fall apart during implementation. Many times data collection requires more time than originally anticipated. Often the protocol must be revised as problems occur. When planning for the study, make sure you consider how data will be collected, what forms will be used to record the data, and who will be responsible for it.

Analyze the Data

Once the data collection phase is completed, the data are summarized in tables and graphs using basic descriptive statistics. If the study design requires it, formal statistical procedures are used to test hypotheses. Finally, you must interpret the findings and form your conclusions.

Publish the Findings

There is no point in doing all the work of a study if you do not communicate your findings to your colleagues. And you cannot effectively communicate them unless you write a report. Of course, once as long as you are going to write them anyway, you might as well use a style recommended by one of the medical journals. The manuscript can be a simple one page abstract. You can then submit the abstract to the journal for review and possible publication. If it is published, it will be preserved as part of medical history in copies of the journal world wide.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- Hypothesis
- Rejection criteria

True or False

1. Experiments are designed to prove that a hypothesis is true.
2. The following is the correct series of steps for using the scientific method:
    a. Formulate the problem
    b. Generate a hypothesis
    c. Define the rejection criteria
    d. Perform the experiment
    e. Test the hypothesis

Multiple Choice

1. All of the following are steps in conducting scientific research except:
    a. Search the literature
    b. Design the experiment
    c. Analyze the data
    d. Survey patients
    e. Publish the results
2. Comparing the experimental results to the rejection criteria is a part of which step in the scientific method:
    a. Creating the problem statement
    b. Formulating the hypothesis
    c. Creating the prediction
    d. Formulating a conclusion
    e. Testing the hypothesis

Chapter 5. Developing the Study Idea

People often say that emotion and personal belief play no part in the scientific method; that only through detached objectivity can the truth be revealed. If this were in fact the case, there would be no human scientists. Without passion, there could be no hypothesis; without a hypothesis there could be no experiment; and without experimentation there would be no science. This chapter examines the factors that motivate and guide the development of a research protocol.

SOURCES OF RESEARCH IDEAS

Choosing and defining a research topic are the first steps in applying the scientific method to a clinical research problem. This process implies concern or question about some concept or observation. Indeed, the scientific method itself can be viewed as nothing more than organized curiosity. Curiosity about the details of one's everyday activity provides the impetus for finding out how or why events are related. The scientific method simply provides a standardized and efficient technique for describing these relationships in a way that can be verified by other observers. Curiosity and the creative energy it engenders are vital ingredients of a productive research effort.

In choosing a research topic, one's interest may be stimulated in a number of ways. One of the most obvious ways is to read journals such as *Respiratory Care, American Journal of Respiratory and Critical Care Medicine, Critical Care Medicine,* and others that are devoted to cardiopulmonary medicine to discover what other researchers are doing. Often one investigator's results will not completely answer the questions another investigator seeks to answer. Perhaps the authors themselves suggest areas where further work needs to be done. Occasionally, the results of an article contradict those of a previous study, creating the need for yet another look at the research problem. Review articles that cover the state of the art in some area of research are especially fruitful in generating ideas along these lines.

The basic concept to remember is that research breeds more research. With this idea in mind, another source of research problems can be identified. We can start with a well-established theory or concept of patient care and test whether or not it is true for a particular set of circumstances. For example, the theory that positive end-expiratory pressure (PEEP) helps to reestablish functional residual capacity (FRC) and thus improves lung compliance is well recognized. Is this theory true for all patients? What about premature infants with respiratory distress syndrome? Perhaps something about the immature lung causes it to respond differently to PEEP than the adult lung.

Another example is in the area of pulmonary function testing. Several common bedside pulmonary function evaluations, such as spontaneous tidal volume, maximum inspiratory pressure, and dead space/tidal volume ratio, are frequently recommended for evaluating patient readiness for weaning from mechanical ventilation. How well do these criteria predict successful weaning? Are other indices, such as the respiratory quotient or urine output, better predictors? Much pressure is imposed on health care departments to justify procedures because of the increasing difficulty of obtaining reimbursement funds. The time is ripe for many established practices to be reevaluated.

A third major source of research topics is the realm of personal experience Your day-to-day work experience constantly provides opportunities to ask, "Why are things done this way?" or "What would happen if. ..?" Consider how many daily decisions are based on tradition or authority, without any

apparent objective rationales. Often an event that causes a sense of irritation, especially if it is a recurrent event, can be turned into a legitimate research problem that may result in a solution or a better way to do things Talking with co-workers can also be a valuable source of insight and ideas.

Trying to develop research topics from personal experience is often the most frustrating approach for the beginning researcher. The natural tendency is to choose a general problem that everyone seems to recognize but no one does anything about. The difficulty lies in trying to narrow the general idea to a specific problem statement. There are at leas two reasons. First, a general problem, by its nature, is often spoke about in vague, undefined terms. Second, in attempting to explicitly describe, the problem, the investigative task may appear to be overwhelming. One may easily become frustrated to the point of not being able to write anything.

One way to avoid this situation is to start small. Begin with a specific incident that stimulated either curiosity or irritation. Simply state what you see happening and why it is important. Write a narrative, first-person account of the incident. This will help you to begin to formulate the research problem with a feeling of involvement and prevent impulsive generalizations.

DEVELOPING A PROBLEM STATEMENT

Once a specific idea for a research topic has been selected, the next step is to develop a formal problem statement. This problem statement is the foundation of the actual study design. It dictates the concepts and methods used to gather data. It also determines the theoretical context in which the conclusion will be interpreted. Moving from a rather loose set of ideas to a formal problem statement, however, is a rather complex process. It involves many false starts as original notions are discarded or modified after careful consideration.

The development process begins with an expansion of the scope of the original problem. This stage involves an *inductive* reasoning process of going from specific observations to general theories. At this point, the words and concepts used must be explicitly defined. Definitions not only help to organize one's thoughts but also makes it possible to relate the topic to previous research. Explicit definitions also facilitate communication with experts whose experiences are helpful in forming a more realistic perspective of the problem.

Now you begin reviewing the literature, using key definitions to speed the search. Try to find analogous problems in other disciplines to create original experimental approaches. For example, many problems concerning clinical measurement (e.g., airway pressure measurement) have been solved in the context of electrical or mechanical engineering. Another aspect of expanding the problem's scope is the identification of any pertinent theories that might be useful in establishing a relationship among a problem's many elements.

For example, suppose you were originally concerned about the maximum respiratory rates newer mechanical ventilators were capable of delivering. Perhaps no guidelines exist for selecting a maximum frequency for a patient, and you believe that rates as high as 100 to 150 breaths/minute may be inappropriate. Perhaps you have observed a few patients whose condition deteriorated when ventilated with high rates. In the original narrative account of the research problem, the term *gas-trapping* was used. During the process of expanding the original idea, consulting with experts, and reviewing the literature, the concept of gas-trapping is found to relate to other specifically defined terms, such as lung compliance, airway resistance, and functional residual capacity. Note that the definitions used here are *operational* definitions; that is, they are defined in terms of the specific operations, observations, or

measurements used in gathering the data. Now the problem can be stated more explicitly forming some general concepts that seem promising.

Your focus of attention should shift from a review of clinically oriented literature to articles dealing more specifically with pulmonary physiology. You may find that the interaction among ventilator frequency, lung compliance, and airway resistance is sometimes described using the concepts of impedance, time constant, and frequency response. A new idea is discovered: The actions of the various components of the respiratory system can be modeled using electrical analogs. Thus, you are led to review electrical engineering texts for a complete understanding of these concepts. This is an excellent example of how a study of analogous problems in related disciplines can help create a general theory by enlarging the scope of the problem.

Once a generalized theoretical framework has been established relating the various concepts that are associated with the research topic, you must narrow and refine those ideas. This is a *deductive* process of going from general theory to a specific application and experimental design. At this stage you need to formulate a specific hypothesis and describe specific experiments to test it. The experimental plan is a description of the variables that will be measured and how they will be interpreted. Using the previous example, you may decide to measure pulmonary compliance and resistance, cardiac output, and the FRC in a series of ten patients. These measurements will be used to test the hypothesis that patients with high resistance and compliance require lower ventilatory frequencies to avoid inadvertent increases in FRC and concomitant decreases in cardiac output.

A formal statement of the research problem might be: "The optimum ventilator frequency for patients with chronic obstructive pulmonary disease is less than that for patients with normal lungs." Of course, the term *optimum* must be defined in terms of measurable criteria. The problem statement, along with the supporting theoretical framework and experimental design, constitutes the research protocol.

Finally, a formal research protocol must be drafted using the specific outline required by your Institutional Review Board. This outline will include a consent form that must be written in non-technical lay terms.

Keep in mind that not all research topics would be developed in exactly this format. Development of the topic depends on the level of the research. There are three basic levels of research. At the first level, you are primarily interested in describing *what* occurs in a particular situation. Typically, little or no literature is available on the topic and the research design tends to be exploratory or descriptive. The problem statement is often declarative in nature, and the research method centers on specific laboratory or patient observations and questionnaires. The data analysis is usually based on descriptive statistics. Examples of this type of research appear frequently in the "Methods and Devices" section of journals such as *Respiratory Care.*

The second level of research includes a description of the *relationships* between or among variables. Usually there is some literature available on the topic, but not enough to predict the action of the variables. The problem statement may be in the form of a question, and the research method is often a descriptive survey using structured questionnaires or physiologic measurements. This method differs from that of the previous level in that some conclusion can be made about *how* variables are related. The data analysis typically involves correlation and regression statistics. Examples of this type of research can be found in the many studies comparing transcutaneous blood gas values with traditional arterial blood gas analysis or comparing arterial oxygen saturation via pulse oximetry with directly measured oximeter values.

The third level of research seeks to explain *why* variables are related in a particular manner. Usually enough literature is available on the topic to predict the action of specific variables based on a particular theoretical framework. The problem statement should be in the form of a hypothesis, and the research method involves randomized controlled experiments. Data analysis is usually in the form of statistical procedures that distinguish significant differences between estimated population parameters.

JUDGING THE FEASIBILITY OF THE PROJECT

The process of identifying and defining a research topic, because of the personal commitment involved, can be very nearsighted. That is, after all the effort you put into developing a research topic, you may find it difficult to see that others might not be interested in supporting the project or that feasibility problems make the study's implementation questionable. Thus, before beginning the study, step back mentally and evaluate the overall worth of the project in a larger context. The major considerations are listed in Table 5-1.

TABLE 5-1. Factors affecting the feasibility of a research project.

1. Significance or potential benefits of study results
2. Measurability of research variables
3. Duration and timing of study
4. Availability of research subjects
5. Availability of equipment and funds
6. Knowledge and experience of investigators

Significance of the Problem

What are the potential implications of the study results? Will a specific population of patients, the medical community, or society in general benefit from the proposed study? Will the results lead to practical applications or an expansion of medical knowledge? Will the findings support or challenge untested assumptions? These types of questions should be asked in the early stages of the project.

The researcher should examine his or her personal motives. Will the research effort culminate in a published journal article? If one is interested in describing a *breakthrough* discovery in a *classic* article, the research problem should be of the scope that it could be linked to a general theory so that the results will have broad application. On the other hand, problems of a more limited scale can be very relevant to departmental policy-making. Every author enjoys the satisfaction of having a manuscript accepted for publication. The goal of a study, however, should never be the advancement of personal prestige. Efforts should be concentrated on providing the highest quality of research. Fame will take care of itself.

Measurability of the Problem

To use the scientific method to investigate a research topic, the problem statement must involve *variables* that can be precisely *measured* and *defined.* Thus, questions such as those concerning morals or ethics are usually not appropriate research topics. However, such questions may possibly be modified

to allow research of a related aspect of the general issue. For example, the research question, "Do patients have the right to die?" is too arbitrary to be an appropriate topic. Consider an alternative question, "Does a healthcare worker's experience reflect his or her opinion about euthanasia?" This topic may be studied by means of a survey, and the knowledge gained might be useful in developing an understanding of the general ethical issue, which in turn facilitates future decision making.

Time Constraints

A good research plan will specify the expected amount of time necessary to complete the project. This is necessary to evaluate other factors, such as patient availability and cost. In addition, it may be necessary to time the study to coincide with optimum data collection. For example, a study of postoperative cardiac patients should coincide with the surgeons' schedules if there are seasonal fluctuations in the types of patients they see.

Availability of Subjects

Research subjects may be anything from ultrasonic nebulizers to rabbits to human infants. If inanimate objects are to be used, make sure you have an adequate supply of functional units. If the study involves animals, suitable facilities must be available and specific guidelines must be followed to insure their humane care. If the experimental subjects are humans, the cooperation of a physician will be needed.

If the study is of the type that tests a hypothesis, the most crucial aspect of determining subject availability is knowing how many subjects will be needed. You must perform some elementary statistical calculations *before* the data are gathered to ensure the usefulness of the statistical conclusion *after* the data are gathered (see discussion of sample size in the chapter on basic statistics).

Cost and Equipment

In addition to planning for the proper equipment and space to conduct a study, consider the cost involved. If one intends to apply for a research I grant, it is essential that expected costs be itemized. A partial list of expense categories might include:

- Literature costs: Index cards, journals, literature searches, manuscript copies, and illustration fees.
- Labor costs: Reimbursement to subjects for cooperation, salary support for technicians and secretaries, laboratory tests, and consultation fees.
- Supplies and laboratory equipment: Paper, notebooks, recorders, transducers, and amplifiers.
- Transportation costs, if the results will be presented at a national convention.

Experience

Although you may be able to obtain help with the various aspects of the study, you should avoid sophisticated measuring instruments or complex statistical analysis for your first study. Planning the project from beginning to end helps to avoid the demoralizing experience of having to abandon the study because of being "in over your head."

SUMMARY

Selecting a clinical research problem is a nebulous process. Beveridge* has said

> *"It is not possible deliberately to create ideas or to control their creation. When a difficulty stimulates the mind, suggested solutions just automatically spring into the consciousness. The variety and quality of the suggestions are functions of how well prepared our mind is by past experience and education pertinent to the particular problem. What we can do deliberately is to prepare our minds in this way, voluntarily direct our thoughts to a certain problem, hold attention on that problem, and appraise the various suggestions thrown up by the subconscious mind."*

Appraising the various aspects of a given research problem is the process of identifying and defining specific variables and placing these variables into a unifying theoretical framework. This appraisal is essentially an inductive intellectual process, reasoning from specific observations to general theories. What follows is the deductive process of reasoning from these general theories to particular hypotheses, and the creation of experiments to test them. There can be no rigid order to the steps used in formulating a research statement. The general thought process, however, should roughly follow the flowchart illustrated in Figure 5-1 Experience will dictate one's personal style. Remember to include operational definitions, a literature review, and an assessment of feasibility.

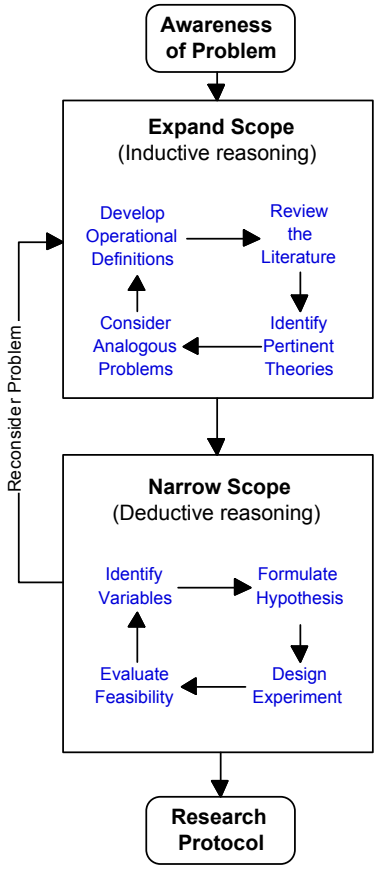


Figure 5-1. General thought process involved in developing a formal research protocol. Beginning with an awareness of a problem, the scope of the topic is expanded by a recursive process of creating definitions, reviewing the literature, identifying relevant theories, and considering similar problems in other areas. A theoretical framework or conceptual model can then be created. Next, specific variables and hypotheses suggested by the model are developed, and experiments designed to test the hypotheses. If the experiments do not seem feasible, reconsider the study problem. You may have developed a new perspective that would lead you to start over using a slightly different line of reasoning.

* Beveridge WID. The art of scientific investigation. New York: Vintage, 1950.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- Inductive reasoning
- Deductive reasoning
- Operational definitions
- Feasibility analysis

True or False

1. A major source of potential research ideas is medical journals.
2. The problem statement is important because it dictates the concepts and methods used to gather data.
3. The general thought process in developing a research protocol is to go from a narrow scope to an expanded scope.

Multiple Choice

1. Deciding if the results of a study may lead to practical applications applies to which factor in the feasibility analysis?
    a. Significance of study results.
    b. Measurability of research variables.
    c. Duration and timing of study.
    d. Availability of research subjects.
    e. Availability of equipment and funds.
    f. Knowledge and experience of investigators.
2. Deciding to consult a statistician applies to which factor in the feasibility analysis?
    a. Significance of study results.
    b. Measurability of research variables.
    c. Duration and timing of study.
    d. Availability of research subjects.
    e. Availability of equipment and funds.
    f. Knowledge and experience of investigators.

3. Realizing that the sample size required will take too long to gather applies to which factor in a feasibility analysis?
    a. Significance of study results.
    b. Measurability of research variables.
    c. Duration and timing of study.
    d. Availability of research subjects.
    e. Availability of equipment and funds.
    f. Knowledge and experience of investigators.
4. Which factor in a feasibility analysis would prompt you to see if rabbits were available in the medical school's animal lab?
    a. Significance of study results.
    b. Measurability of research variables.
    c. Duration and timing of study.
    d. Availability of research subjects.
    e. Availability of equipment and funds.
    f. Knowledge and experience of investigators.

Chapter 6. Reviewing the Literature

Once you have identified a clinical problem, the next step is to review the available literature. A thorough investigation of others' work at this point can save time, money, and trouble. Ample research may already have been done on a particular area of interest. On the other hand, opportunities for further research may be quickly revealed. Regardless of the techniques and sources used, literature review is a vital step in any health care research project.

Why conduct a review? First, the literature review enables you to discover whether related studies have already been done. If there have been six studies showing that the highest obtainable F_IO_2 reached with a particular manual resuscitation bag is greater than 0.90, then another study is obviously not needed to reconfirm these findings. If, however, there has been only one study comparing the aerosol output of two ultrasonic nebulizers, and the results are highly controversial, this could be a useful area to explore. In fact, any area in which there is a great amount of controversy would be one worthy of further exploration. A single study is not conclusive. Only after several investigations have replicated a study's original findings can they be considered valid.

Researchers must immerse themselves in the literature on a specific topic to become knowledgeable about previous related studies. Research on a particular subject should be conducted by those who are well versed in this area and are able to converse intelligently on all aspects of the topic. This has the benefit of allowing communication with other researchers and experts in the field, to "pick their brains" if a problem arises during the course of the research. You will be encouraged by directly with recognized experts in an area of inquiry who are on the cutting edge of knowledge. Although other studies may have been done in a certain area, you may have a completely different approach. Finally, you may have reassess the plan of action. Perhaps the scope of your project should be narrowed or even refocused on another area entirely.

A further reason for performing a literature search is to learn how others designed their studies. Proper design can make the difference between a significant study and one that is useless. If the intent is to replicate earlier work, this investigation will enable use of the earlier author's design, if it is a good one. By looking at several different study designs, you can judge which may be best suited to address the current problem. Perhaps an entirely different design will better serve the purpose. In fact, by noting where previous authors had difficulty, you can avoid duplicating earlier mistakes.

CONDUCTING THE LITERATURE SERACH

Scope of the Review

If the investigation is of a relatively new device or procedure this may be feasible. A search of all available literature on the topic is not only possible but desirable if a limited number of articles exist.

Other areas of inquiry may have a much more extensive base in the literature. In these cases, fiscal and temporal constraints will make an exhaustive search impossible. Therefore, the investigator will have to decide how to limit the inquiry. One means of limiting the number of citations included in a search is to look at only the last three or five years. This will provide the most recent research, some of which may be replications of and thus confirmations or refutations of earlier studies. These recent articles will likely

refer to a few significant earlier works that formed the basis for many subsequent studies. By including a few of those earlier articles, it may be possible to obtain the historical "flavor" of a particular topic. This would be especially important if there were controversy surrounding the topic.

If you have difficulty in narrowing the field to a manageable size, a search of the past year will show the popular topics. This search may also cause an alteration in direction of the proposed study. The topic of interest may have been extensively and conclusively detailed by others. Conversely, there may be an obvious gap in the literature that merits further study.

The importance of performing a proper literature search at this stage cannot be overemphasized. As a first time researcher, you will have many questions about general principles that a more experienced researcher can answer. In fact, identifying a mentor to provide guidance throughout the project is essential if you want to successfully complete your first project. This mentor can provide assistance in obtaining funding, serve as a sounding board when difficult decisions need to made, and identify other resource persons to assist in specific areas, such as statistical analysis.

Performing the Search

It is important from the start to have a system for filing and cataloging data before beginning to read even the first article. Having a plan for dealing with a large amount of information will save time, energy, and frustration when it comes time to write a concise study report.

One example of such an information management system is the use of 3 by 5-inch file cards (or if you prefer, a personal digital assistant such as a Palm computer). The first part of the entry should be a complete, accurate citation, such as would appear in the bibliography at the end of the paper. Be sure to record this before doing anything else. If the article is obtained through inter-library loan, the computer printout that accompanies the article should be saved, as it may contain the only complete journal citation.

Accuracy of these citations is importance and, as stated in many journals, is the sole responsibility of the author. If readers have further interest in the topic, or question the results of the research, they might return to cited articles to recheck their impressions. If the citation is inaccurate, the reader is in a perplexing situation. This error may also cast further doubt on the credibility of the study. If an author cannot keep his or her citations straight, the reader is likely to lose faith in the results of the research. One of the most boring aspects of having an article published is proofreading both the copy before it is typeset and the galley proofs. The references demand the closest inspection. The editor will check only for proper form and punctuation. The author must scrupulously check, number by number, symbol by symbol, to insure there are no mistakes.

Once the citation is recorded accurately, writing a synopsis is the next step. Although brief, the synopsis should contain all-important points made in the article. It is a good idea to underline or highlight points that are important or perhaps do not make sense. Once the entry is complete, it should be double checked against the original journal article. This practice may seem an unnecessary duplication of effort but it is certainly worth the time. Many people are likely to read the completed research conclusions. Should there be an error in the use of sources, this will likely be detected. If the writer is fortunate, a knowledgeable reviewer will pick up the error before the article is accepted for publication; then the error will cause only minor embarrassment. If, however, the faulty assumption was one of the bases for the research, and the article is published, this could be a personal and professional calamity

This is an opportune time to mention the use of a personal computer, microcomputer, for recording and organizing the information from a literature search. A microcomputer with word processing or database

management software will save time and effort in managing large amounts of information. Always maintain a backup disk in case there is a problem. This backup should be updated after every session and should be stored at a site that is remote from the master copy. It is also a good idea to maintain a printout as backup. Another important technique to remember is frequent saving of files onto the disk during a work session. Normally, all user programs are stored in random access memory (RAM) during use. Data can be read from or written into RAM, but are not retained when the power is turned off. If there were a power failure from an electrical storm or from someone's tripping over the power cord, all data in RAM would be lost. Thus, the prudent operator saves data in the disk approximately every half hour. Don't fall victim to the old saying "There is never enough time to do it right, but there is always time to do it over."

SOURCES OF INFORMATION

Books

The student or novice researcher may think that textbooks would be the first place to look for information about the research problem. The problem with books is that the information they contain is usually outdated before it is even published. Consider that it takes between 2 and 5 years from the time an author starts writing to the time the book appears on store shelves. Consider further that the reference papers the book is based on take 1 to 2 years to get published and most of the papers referenced are not new. So, at best, the information in textbooks is 3 to 7 years old and most of it is much older than that.

On the other hand, textbooks are usually a good place to look for basic theories and concepts that you can use as a foundation for building your own hypotheses. They are also good for learning about measurement, statistical, and even patient care techniques needed for designing your experiments. Finally, textbooks often have good lists of reference articles that can be a start in further literature searches. Pay particular attention to the names of authors who seem to be authorities in particular subjects. Then you can search for more recent articles by these authors.

Journal Articles

Most of the information you will use in formulating your research plan will come from medical journal articles. The question is how to find the right information. You must first understand that for just about any topic that has been researched you will find conflicting results in published papers. That means you must try to read at least the abstracts of all papers published in the last 3-5 years on your topic, if such papers exist. Then, you must identify the controversial areas and those that are relatively settled. The controversial areas will provide ideas for new research and the settled areas provide ideas for experimental designs.

When searching for relevant articles, pay particular attention to recent papers that are reviews of existing knowledge. For example, you might find an article entitled "Alternative to percussion and postural drainage: A review of mucus clearing therapies." Such an article will save you a lot of work because the author has already searched the literature to find and summarize the published articles on that particular topic. Not only will the review itself be helpful, but also its list of references will be useful for further searches.

A growing number of periodicals summarize the best evidence in traditional journals. They provide structured abstracts of the best studies and expert commentaries. These new journals include *Evidence-Based Medicine, Evidence-Based Nursing,* and *Evidence-Based Cardiovascular Medicine.* These

journals do what we can't do for ourselves; summarize the best evidence from high-quality studies selected from all the journals of relevance to our research interests.

The Internet

Online searches using a personal computer with an Internet connection can put a world of information at your fingertips. If you don't have Internet access yourself, most libraries offer free service. In addition, the librarian will help you learn how to conduct searches.

The National Library of Medicine offers free searches. Links to other related databases and online journals can be accessed from this site. Abstracts are easily obtained in addition to a listing of articles on the subject matter and can be printed or saved to a disk. Some journal listings have links to full text articles available online. Table 6-1 lists a number of Internet sites that you will find useful. Most of these sites give instructions for creating searches.

Table 6-1. Useful Internet sites

Database	Web Site	Content
MEDLINE	www.ncbi.nlm.nih.gov/PubMed	Worlds largest general biomedical research literature database
Cochrane Library	www.update-software.com	Systematic reviews
ACP Journal Club	www.acponline.org	Synopses of articles for internal and general medicine
Centers for Disease Control	ww.cdc.gov	Health information
VentWorld	www.ventworld.com	Everything related to mechanical ventilation
The Merk Manual	http://www.merckmedicus.com/pp/us/hcp/hcp_home.jsp	Research diseases Patient resources Technology resources Drug reference Harrison's online Merk Manual online
CINAHL	www.cinahl.com	Cumulative Index to Nursing and Allied Health Literature

HOW TO READ A RESEARCH ARTICLE

Regardless of where one obtains information, the source must contain certain basic components, and these components must meet certain criteria. Whenever you read a published report, you must maintain an open mind as well as a healthy skepticism. You should ask a number of critical questions:

1. Does this information make sense?
2. Does what the author is saying consistent with what I know to be true?
3. If not, does the author make a convincing case?
4. How good are the author's methods?

The following is a brief description of the elements that should be contained in a research paper, and what questions to ask when judging each section. Most research studies have the same basic format: *Abstract, Introduction, Methods, Results Discussion,* and *Conclusion.*

The Abstract

The abstract is a condensed version of the paper, usually containing brief versions of the introduction, methods, results and conclusion. Online searches will generally result in a list of titles that are linked to the papers' abstracts. Many times, you need only read the abstract to get the information you need. Print the abstracts of the papers relevant to your project during your literature search. Use them later for reference when you go to the library to find the full articles.

The Introduction

The purpose of the introduction is to inform the reader about what was studied and why. It will generally provide a brief literature review on the research subject to justify the importance of the study. Most importantly, the introduction should contain a clearly stated research question, or hypothesis. For example, in one paper submitted for review, the author stated that the purpose of the study was "to investigate the output of nebulizer X." This statement is not specific enough. A better statement might have been "to determine mass median particle size, aerosol density at different flow rates, and volume of water aerosolized per unit time." The first statement leaves the reader to guess exactly what about the nebulizer is to be studied. After reading such a study, you cannot tell whether the author accomplished the objective, since none was clearly stated.

There are three types of hypotheses: descriptive, correlational, and comparative. A descriptive hypothesis asks how, what, or how much and is not statistically testable. A correlational hypothesis seeks to determine the type of relationship between two variables. A comparative hypothesis seeks to determine the difference between X and Y, as in treatment versus no treatment. Both correlational and comparative hypotheses are appropriately tested statistically.

The Methods Section

The purpose of the methods section is to describe the experiment(s) in enough detail that other researchers are able to reproduce the study. If others are not able to obtain the same results using the same methods, the hypothesis will eventually be rejected, no matter what the original results were.

First, determine whether the experimental subjects (pieces of equipment, animals, or people) were comparable to those you wish to study. The idea is to ascertain if the conclusions reached by the authors are relevant to the subjects you wish to study.

Next you need to examine the outcome variables. These are the variables measured to test the hypothesis. For example, if the introduction stated that the purpose of the study was to evaluate the accuracy of a new blood gas analyzer, you would expect to see blood gas measurements (e.g., PaO_2, pH, etc). But you would also expect some calculated *differences* between measured values and known (or expected) values (such as those from tonometered blood; blood equilibrated with known concentrations of oxygen an carbon dioxide). Such calculated outcome variables are used to assess accuracy. If the introduction stated the hypothesis was "removing the inner cannula of a tracheostomy tube decreases the work of breathing" then you would expect to find outcome variables related to measurements of pressure, volume, and flow plus calculated values of work. When calculated outcome variables are used, make sure you understand the math if you intend to use similar variables. Determine if the authors have provided enough outcome variables to test all the hypotheses mentioned in the introduction and avoided any variables that are not relevant. When doing background research, you are most often looking for how things were measured, including how the measuring devices were calibrated.

The methods section will give you ideas of the type of study design that is most appropriate for your topic, even if you don't yet know much about study designs. You will learn by example the differences among randomized controlled trials, case studies, device evaluations, and retrospective reviews. Table 6-2 lists some terminology associated with research design.

Table 6-2 Terms used to describe features of research studies.

Term	Meaning
paired comparison	Subjects receiving different treatments are matched for confounding variables such as age and gender. Results are analyzed in terms of differences between subject pairs.
unpaired comparison	Each group of similar subjects receives a different treatment. Results are analyzed in terms of differences between groups.
crossover	Each subject received both the intervention and control treatments (in random order)
randomization	The scheme by which subjects and treatments are paired so that each subject has the same chance to receive each treatment. The purpose is to make sure the subjects in different treatment groups are as similar as possible.
single blind	Subjects did not know which treatments they were receiving.
double blind	Neither investigators nor subjects knew who was receiving which treatment.
placebo controlled	Control subjects receive a placebo (inactive treatment) which should appear the same as the active intervention treatment. Placebo (sham) operations may also be used in trials of surgical procedures.

Finally, the methods section will also describe the statistical methods used to describe the raw data and to test the hypotheses. There are hundreds of different kinds of statistical procedures. Only a handful is used in medicine and even fewer in the health care literature. By examining the statistical methods you will not only find out what procedures are common for the type of research you want to do but in many cases how to do it or at least references that will tell you how to do it. You will also learn valuable information about specific types of equipment and brand names that are commonly used in research.

The Results Section

The first thing to do in the results section is to check all the outcome variables mentioned in the methods section (and implied in the introduction's hypothesis or problem statement). The big lesson to be learned in this section is how to present data in tables and figures. Take note of the kind of graphs used and how tables are designed.

The other major item in the results section is whether the hypothesis tests (if any) yielded significant results. Sometimes the results are not statistically significant only because the sample size was too small (which you could check again in the methods section). The concepts of statistical versus clinical significance and sample size will be discussed in the chapter on statistics. By the time you are finished with the results section, you should know if the study question from the introduction was adequately addressed.

The Discussion

The purpose of the discussion section is to provide a more detailed review of the background of the study and to link the data (results section) to the hypothesis (the introduction) and explain any problems or limitations (connecting with the methods section).

This section allows the author(s) to expand on the background and justification for the research problem. Often many references will be cited. Take the time to look at the titles of the references when they are cited in the text. Make sure they are both relevant and recent. Note any that you would like to obtain from the library. The authors should discuss how their findings agree or disagree with those of previous studies.

Often you will find detailed discussions of any theoretical concepts related to the research problem that you can use in formulating hypotheses and experimental designs for your study. The authors may also discuss problems or limitations with their study and suggestions for other researchers.

Conclusion

The conclusions are often stated either in a separate section. Here is where the authors try to convince you that the data they collected answered the study question and provided new scientific insight. In a good paper, you will learn valuable lesions in logic and scientific writing. Scientific thinking is something you learn by example. Following the example of good authors and mentors is the best way to learn.

Summary

The ability to critically examine journal articles, letters, editorials, and research papers is a skill that all healthcare practitioners should possess. During the course of discussions with other health care specialists and physicians, your opinions are valid only if they can be supported by objective research findings. In conducting original research, you must be able to evaluate what has already been studied.

Finding the information in the first place has recently become more convenient with the proliferation of databases and companies that specialize in data retrieval. Although the speed of computers makes the drudgework tolerable, the machines are incapable of telling the investigator what to look for. The investigators themselves, who can reason and judge, are still the most important part of the system.

QUESTIONS

True or False

1. If your study idea has already been published, you should abandon it.
2. A literature review is necessary only if you do not know anything about your study problem.
3. One good reason to perform the literature review is to see how other researchers designed similar studies.
4. Three good sources of information for a literature review are books, journal articles, and the Internet.
5. Books are the best source of information because they are themselves based on literature reviews.

Multiple Choice

1. In which section of a published research article would you find a brief overview of the study?
    a. Abstract
    b. Introduction
    c. Methods
    d. Results
    e. Discussion/Conclusion
2. In which section would you find a description of the statistical analysis?
    a. Abstract
    b. Introduction
    c. Methods
    d. Results
    e. Discussion/Conclusion
3. In which section would you expect to find the authors' interpretation of the experimental data?
    a. Abstract
    b. Introduction
    c. Methods
    d. Results
    e. Discussion/Conclusion
4. Where would you find a statement about the research problem or a hypothesis?
    a. Abstract
    b. Introduction
    c. Methods
    d. Results
    e. Discussion/Conclusion
5. In which section would you find any *p* values associated with statistical tests?
    a. Abstract
    b. Introduction
    c. Methods
    d. Results
    e. Discussion/Conclusion

Chapter 7. Designing the Experiment

Sampling of subjects or units for observations and choice of an appropriate research design are two links in the logical chain of the research process. That process began with the definition and delimitation of a problem, a review of the literature (prior knowledge on the subject), identification of the research variables, and the translation of the research problem into a hypothesis. At this point, the investigator must specify the population of interest, decide whether and how to choose a sample, and choose a research design appropriate to the research question.

SAMPLES AND POPULATIONS

The following definitions of terms will be used frequently in this chapter: Population. The entire collection of cases as defined by a set of criteria

Accessible population: The collection of cases as defined by the criteria and which are available to the investigator.

Target population: The entire collection of cases to which research results (observations or conclusions) are intended to be generalized.

Sample: A subset of the population .

The rationale for making observations on a sample instead of the entire population is conservation of time and money. If a sample is in fact representative of the population from which it is drawn, then measurements from the sample can be validly generalized to the whole population. On the other hand, poor sampling methods can cause sample bias, and the sample will not represent the population. The classically cited case is the presidential voting poll of the *Literary Digest* for the 1936 election. A sample of 12 million persons was selected from telephone directories and automobile registration lists. Of the 21% who returned the voting preference card (Republican or Democratic preference), 57% indicated they would vote for the Republican candidate, Landon. As events turned out, the Democrat Roosevelt was elected by a landslide. Sample size did not make up for the selection bias inherent in the sampling technique. The bias was caused by selecting the sample from telephone directories and automobile lists, which were not representative of the voting population in 1936. The survey mechanism used led to a further self-selection bias, known as *volunteerism:* The 21% returning the survey cards may not have been typical of the entire group receiving cards, because volunteers are not necessarily representative of a whole group.

The example of the *Literary Digest* sampling method underscores the difference between the accessible and target population. If we consider the accessible population to be those on automobiles lists and telephone directories, then the target population was intended to be the general voting public. In fact, the accessible population differed from the target population to which the *Literary Digest* wished to generalize its findings The actual target population was those who possessed cars and telephones, which in 1936, was not the entire voting population.

The distinction of accessible and target populations clearly applies to medical research. Suppose pulmonary function measures are performed in a research study on a sample of in-patient asthmatics, of a certain age and with certain airway reactivity, selected at a large metropolitan public hospital. The accessible population is the group of defined asthmatics at the public hospital. These patients may not be typical of all such asthmatics, since nutritional status and compliance with prescribed treatment often

varies with income and education level. Also, a hospitalized group is usually more severe, and a university hospital, if this is such, may see more complex and serious cases. Do the results on the indicated accessible population generalize to out-patients, to private practices, or to different socioeconomic levels. Finding an accessible population, representative of the population of interest, is frequently difficult.

The following points and recommendations apply to selecting samples for a research study:

- Populations are defined by the researcher, and should be clearly specified. For instance all asthmatics with a positive reaction to some standard antigen challenge, and under the age of 12 years, constitute an identifiable group.
- Populations need not be large in number. For example, all Eskimos who inhale smoked tobacco comprise a small population.
- The accessible population from which a sample is drawn should be clearly described by a researcher in publishing a study, since the accessible population determines the true target population.
- Sample size does affect the precision of estimates, given a certain magnitude of treatment effect, and formulas exist for estimating sample size needed to achieve certain risks of error and precision. With correct sampling techniques, however, probabilities for correct research conclusions can be obtained with very small sample sizes.

METHODS OF OBTAINING A SAMPLE

There are two general classes of sampling: non-probability and probability sampling.

Non-probability sampling occurs when nonrandom methods are used to select a sample from the accessible population. For example, the first 20 patients having arterial blood gas measurements are selected for a study of heparin's effect on measured PCO2. This is a sample of convenience or accidental sample.

Probability sampling is based on random selection, so that a random sample is obtained. Random sampling allows the researcher to specify the probability that each unit in the defined population will be chosen. Because of that fact, the probability of values obtained from the sample can be known. For example, if a sample mean is 10, we can know the probability that the population value is between 8 and 12, if a form of random sampling is used. Thus, probability sampling is essential to the use of inferential statistics in testing hypotheses. There are four types of random sampling methods usually distinguished.

Simple random sampling occurs when every unit in the population has an equal and independent chance of being selected. This is achieved if every unit in the population is numbered, and then a sample is selected by using a table of random numbers. The list of numbered units is the *sampling frame.* A small example illustrates the selection of a simple random sample, using the random digits in Table 7-1. Let the accessible population be the frame in Table 7-1, giving the list of patients' initials numbered 1 to 20. To obtain a simple random sample of 5 from the population, enter the table of random numbers blindly and sequentially choose the first 5 numbers between 1 and 20. For instance, beginning with 08, in row 3, column 2, the first 5 non-repeating random digits between 1 and 20 are 08, 14, 13, 15, and 18. The selected sample is now indicated by bold type in the table.

Table 7-1. Selection of a simple random sample (bold numbers) using a small set of random digits (from a larger table in a statistics textbook or from a computer routine).

Random Digits

20	15	03	10	26	05
13	06	32	12	11	02
12	**08**	**14**	**13**	21	14
15	**18**	09	20	09	04
22	03	07	05	11	20
18	12	15	28	24	17

Sampling Frame

1. J.W	5. J.S.	9. A.R.	**13. V.D.**	17. B.H.
2. A.C.	6. T.S.	10. M.W.	**14. R.D.**	**18. J.T.**
3. D.X.	7. L.W.	11. H.D.	**15. J.R.**	19. J.E.
4. B.T	**8. E.D.**	12. W.C.	16. D.M.	20. J.Y.

Stratified sampling is useful when the population is or needs to be subdivided into strata. A sample is then selected by randomly choosing a specified number from each stratum, using a method such as that described for simple random sampling. A stratified sample can preserve the proportions found in strata of the population, a procedure known as *proportional* stratified sampling.

Systematic sampling is a procedure that, despite its name, ensures a random sample. Briefly, let the population size be N, and the desired sample size *n.* The sampling interval, *K,* is simply *N/n.* The first unit is randomly chosen between 1 and *K,* and then every Kth subject or unit is chosen until a sample of size *n* is obtained. For example, if the population size is 50, and a sample of size 5 is desired, then $K = N/n = 50/5 = 10$. A random number between 1 and 10 is chosen, such as 6. Then units 6, 16, 26, 36, and 46 are selected from the sampling frame.

Cluster sampling is necessary when sampling units occur in intact groups or clusters. For example, if the sampling frame is made up of allergists or practices with asthmatic patients, a cluster sample of these asthmatics is obtained by randomly selecting the desired number of clusters (the allergist, with practice), and then incorporating *all* asthmatic patients in each practice chosen to enter the study.

The list of potential research subjects for a sample usually specifies the accessible population. Be aware that a narrowly defined accessible population limits the ability to generalize results to an equally narrowly defined target population.

When evaluating treatment techniques, mixing populations with no plan can blur results or give false results. For instance, intermittent positive pressure breathing (IPPB) treatments may have different results in a population of abdominal surgery postoperative patients with no history of lung pathology versus a population of diagnosed emphysema subjects. If we wish to study both groups in one investigation, then stratified random sampling would be useful.

Finally, sampling from a population is the basis for inferential statistics. We *infer* from the statistic value in the sample to the population value (parameter). We expect that a sample statistic will differ from the actual population parameter due to sampling error. Knowing the probability of a sample, since it is randomly drawn, allows us to know the probability of any difference between the sample and population values. The following mnemonic helps to relate the terms *statistic* and *parameter:* sample is to population as statistic is to parameter. A statistic is a measure from a sample, and a parameter is a measure from an entire population.

BASIC CONCEPTS OF RESEARCH DESIGN

The research design is the plan or organization for manipulating, observing, and controlling variables in a research question. In this chapter, such plans will be diagrammatically presented using a simple notation for convenience and conciseness. The terms involved are defined as follows:

Variable: An entity that can take on different values. Examples are height, blood pressure, pH.

Independent variable: Variable that is manipulated; the treatment.

Dependent variable: Variable that is measured; the outcome variable of the treatment.

Nuisance variables: Extraneous (usually uncontrollable) variables, also called confounding variables, that can affect the dependent variable.

Placebo: A treatment designed to appear exactly like a comparison treatment, but which has no active component; a presumably inert substance or process to allow a comparison control. In the simplest situation, the research problem is to decide if a change in *X*, the independent variable, causes a corresponding change in *Y,* the dependent variable. Is there a relationship between *X* and *Y*? In most clinical research, this question is not easy to answer because many other factors, termed nuisance variables, may affect the dependent variable. Age, for instance influences pulmonary function measures.

In a two-group design, inherent group differences can affect outcomes. For example, more intelligent health science students may do well with computer-based instruction while slower students may not. Are results due to the instruction or to inherent aptitude? Perhaps the treatment group is consciously or unconsciously handled or evaluated differently, causing more motivation for them to do well. Of course, the change may be caused by the treatment itself, and the lack of change may mean a treatment is ineffective. The advantage of a probability sample is that we can quantify the probability of sampling error that is, the probability of random fluctuation in the measured, dependent variable. For instance, with random sampling we can determine the probability that an observed difference between a treatment and control group is due to chance. If the probability that the effect is due to chance is low, then we may conclude that the difference is due to the treatment and not caused by sampling or random error.

The major challenge dealt with by research design is that of *control.* John Stuart Mill, in *System of Logic,* stated the Law of the Single Variable: If two situations are equal except for one factor, any difference between the situations is attributable to that factor. In the life sciences, holding all factors constant except for the factor (treatment) under investigation is practically impossible. Instead, randomization and probability theory replace the control of absolute equality, at least in the strongest

research designs. There are generally two categories of research designs, experimental and non-experimental.

EXPERIMENTAL DESIGNS

Three characteristics are seen in a scientific experiment:

1. *Manipulation* of an independent (treatment) variable.
2. *Control* of all other variables except for the dependent (outcome) variable.
3. *Observation* of the change, if any, in the dependent variable.

A research design that plans for manipulation, observation, and control is thus an experimental research design, that is, a plan for a scientific experiment. This will not be the case in non-experimental research design, which is considered subsequently.

The major purpose of a research design, especially in clinical research, is *control* of potential nuisance variables. There are four methods of control commonly used in experimental research design:

1. Random selection of sample and random assignment to groups
2. Matching of subjects between groups or grouping of subjects based on a nuisance variable to achieve homogeneity (e.g. grouping based on age or weight)
3. Including a nuisance variable as a treatment variable
4. Statistical removal of a nuisance variable through analysis of covariance .

The first three methods are of *experimental* control, whereas the last is a *statistical* control. The advantage of the first method, randomization, is that random or chance assignment can be expected to even out any nuisance variables far all groups, whether a nuisance variable is known in advance or not. The second method of control is frequently seen when subjects are used as their own controls in a before and after study, or in studies of paired twins. A blocking design, termed a randomized block, will be illustrated when presenting common designs.

Experimental research designs are distinguished from the weakest to the strongest, on the basis of the amount of control employed, using the following terms:

Pre-experimenta. There is little or no control of extraneous nuisance variables. Such a design is often useful for a pilot study.

Quasi-experimenta:. Designs lack full control of all variables, but there is an effort to compensate with other controls. Usually, randomization is lacking, perhaps because of ethical constraints in choosing or assigning subjects to treatment.

True experimental: Designs provide full control of variables by one or more of the methods previously described.

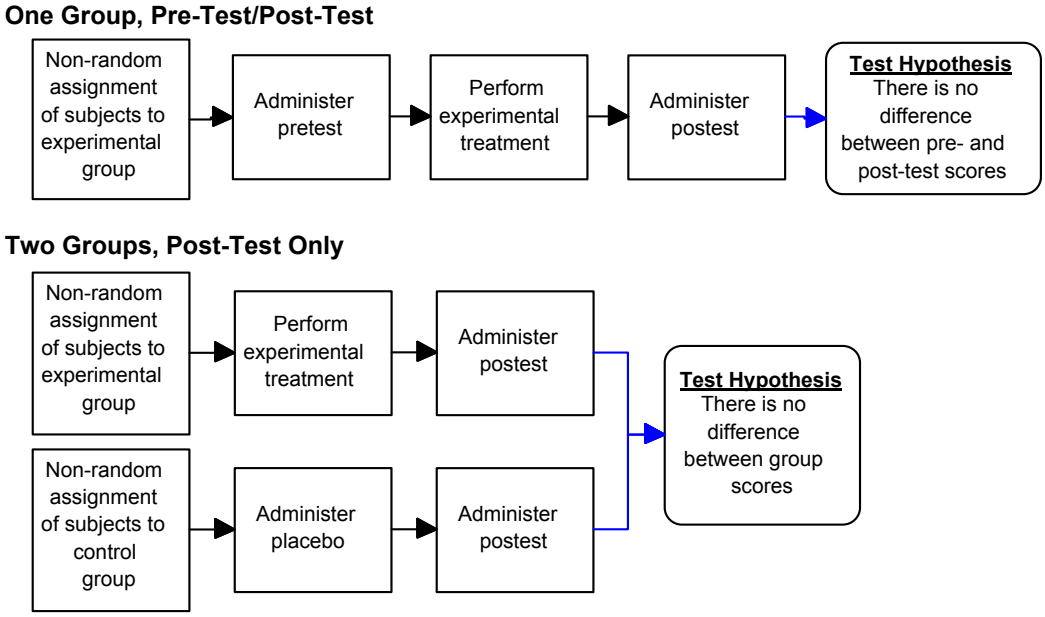
The differences among these three classes of design will be clarified with specific examples.

Pre-Experimental Designs

In a single group, pre-experimental design, the outcome variable is measured, the treatment is given, and then the outcome variable is measured again to see if any change occurred. This is sometimes called the

pretest – posttest design. If there are two groups, no pretest measurement is conducted. The control group receives placebo treatment, and the outcome is then measured (Figure 7.1).

Figure 7.1. Pre-experimental designs.



These designs are considered weak because of poor control of nuisance variables. In the one-group design, we cannot conclude that a change in the treatment caused a change in outcome. Would the same result be seen with placebo? Perhaps the change is simply because the study subjects know they are being studied (known as the "Hawthorn Affect"). The weakness is that no comparison to a group *without* the treatment is available. In the two-group case, if a difference exists in outcome with and without the experimental treatment, how do we know that the difference was not there *before* the treatment was given (inherent group differences)? In both designs seen in Figure 7-1, no random assignment is present to guarantee equivalence of the two groups in the second case and representativeness of the population in the first. However, such designs can be very useful for initial or "pilot" studies.

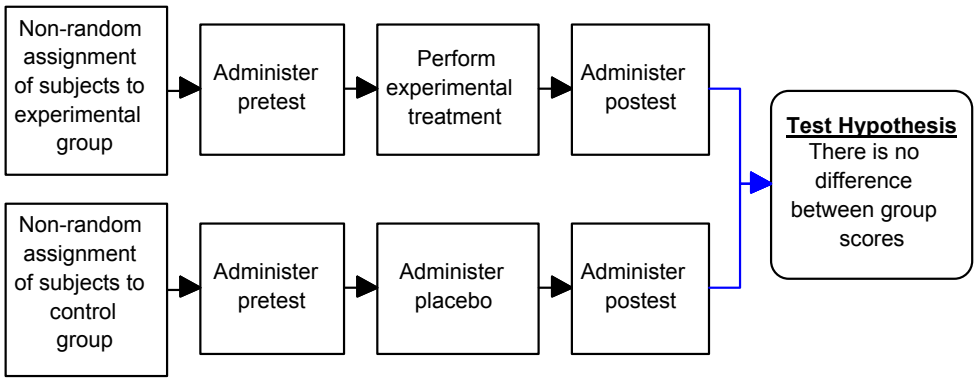
Quasi-Experimental Designs (Case Control)

Case control designs are often used in retrospective studies. *Retrospective studies* are those in which all the events of interest occur prior to the onset of the study and findings are based on past records. Comparisons are made between individuals with a particular condition (the cases) and individuals without the condition (the controls). In order to make cases and controls more comparable, they are usually matched on characteristics known to be strongly related to the condition of interest such as age, gender, race and socioeconomic status.

In the design in Figure 7-2, groups have not been randomly assigned, but there is compensation for the lack of control usually achieved with randomization. There is a comparison group and subjects are measured before treatment or placebo. Any inherent group difference with regard to the dependent variable can be detected with the pre-test, and equivalence of the experimental and control groups can be verified before any manipulation. Although such a design is stronger than the pre-experimental designs

considered, the lack of randomization causes this design to be classed as quasi-experimental. There is no true control of potential differences. For example, suppose drug X and blood pressure were independent and dependent variables, respectively. The pre-test may show there is no difference between the two groups before the treatment, which is important. But the experimental group may be older and respond differently to the drug than younger subjects. Therefore, a difference may be seen *after* treatment that is partially due to age, or perhaps no difference may be seen if older subjects do not respond to the drug. Although the drug may be effective for younger people, it could be described as ineffective, and the pretest would not necessarily reveal this possibility. As we will see in the next section, random assignment of subjects to the two groups would have caused such differences as age to cancel out. However, the design in Figure 7-2 will be useful when intact groups prevent random assignment. Subsequent designs will illustrate how a nuisance factor such as age can be blocked out, or incorporated as a separate treatment variable.

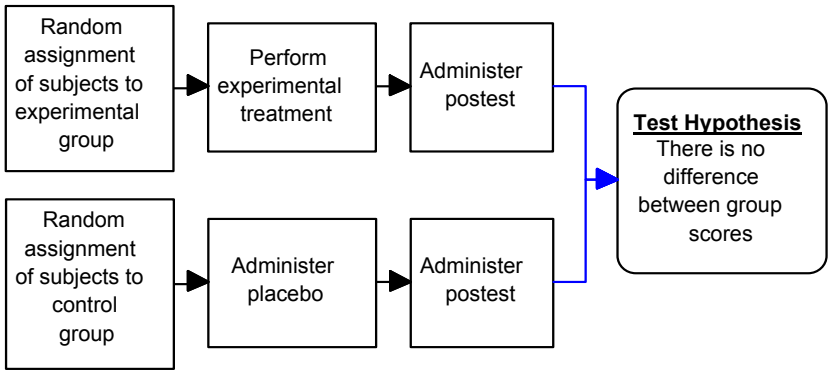
Figure 7-2. Quasi-experimental design (case control).



True Experimental Designs (Randomized Control)

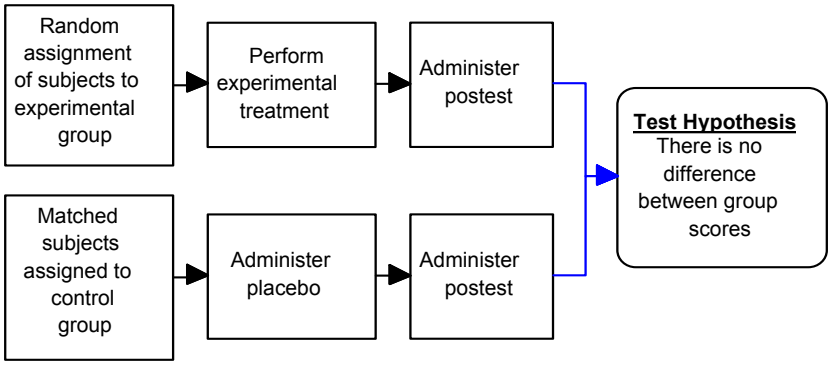
The greatest degree of control occurs with true experimental designs. In the randomized, posttest-only design, any inherent differences such as age, sex, or weight that could affect the dependent variable are presumed to be averaged out between the experimental and control groups. If the two groups are made equivalent by randomization, then a pretest is not needed to determine equivalence (Figure 7-3).

Figure 7-3. True experimental design (randomized control).



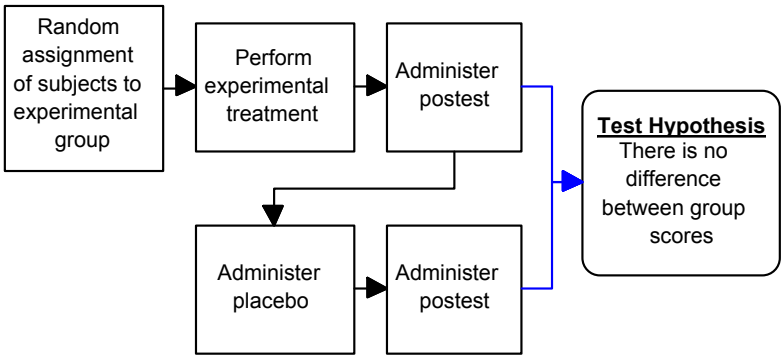
The same design can be modified to match, or pair, the subjects in the two groups (Figure 7-4). The extreme example is a study using twins, where a twin in the experimental group is matched to a twin in the control group. Other forms of matching include pairing of subjects on suspected or known nuisance variables. For instance, a subject in the 30 to 40 year age range in the experimental group has a corresponding, matched subject in the control group.

Figure 7-4. Modified true experimental design (matched control subjects).



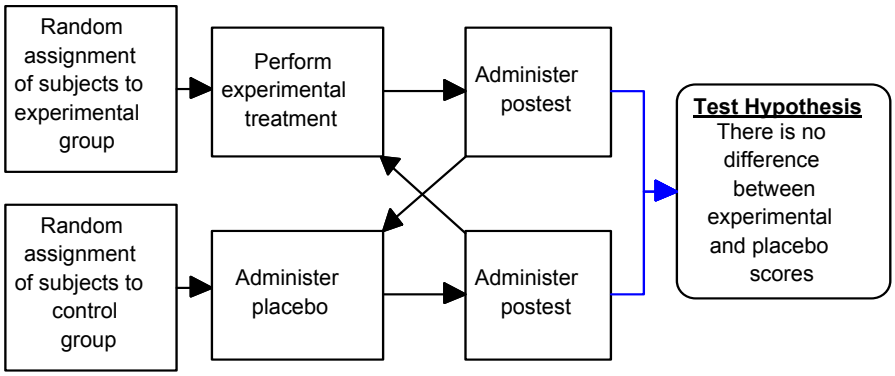
The matched subjects design also includes the case where subjects are used as their own controls. The entire group of subjects first receives placebo, and the dependent variable is measured. Then, the entire group receives treatment, and again the dependent variable is measured. Essentially subjects are matched to themselves (Figure 7-5A).

Figure 7-5A. Modified true experimental design (subjects are their own controls; used only if the order of treatment is not important).



To guard against any residual or learning effect from the order of the treatment and placebo sequence, a *crossover* design may be used (Figure 7-5B). Half of the group receives treatment, then placebo, while the other half receives the reverse order. The entire group still receives the treatment and the placebo, and is measured on the dependent variable.

Figure 7-5B. Modified true experimental design (subjects cross over: used when the order of treatment may affect outcome).



The type of research design can affect which statistical test is used for making a conclusion regarding the research hypothesis. For instance, with two matched groups (as in Figures 7-4 and 7-5), a paired *t* test is required. If the groups are different (i.e., not paired as in Figures 7-2 and 7-3) then an unpaired *t* test is used.

Analysis of variance (ANOVA)

ANOVA is a general statistical technique that applies to a number of experimental designs. A *completely randomized one-way ANOVA* is diagrammed in Figure 7-6. The advantage of this design over those considered previously is that more than two groups can be used. For example, instead of comparing an experimental treatment to placebo, we can investigate three or more dosage levels of the experimental treatment, the independent variable, by using three or more groups of subjects in the experiment. Figure 7-7 illustrates this with the example of the effects of different levels of positive end expiratory pressure (PEEP) during mechanical ventilation on patients' arterial oxygen tension. Subjects are different in different groups, and are randomly assigned to groups. Differences among the groups can be detected with a single statistical test, the F-test (discussed later). In the design in Figure 7-6, there is only *one* independent variable with *n* levels, and there is only one dependent variable being measured.

Figure 7-6, One-way (or one factor) analysis of variance (ANOVA).

Experimental Factor (independent variable)			
Level 1	**Level 2**	**Level 3**	**Level n**
subject 1	subject 1	subject 1	subject 1
subject 2	subject 2	subject 2	subject 2
…	…	…	…
subject n	subject n	subject n	subject n

Figure 7-7. Example of one-way ANOVA comparing arterial oxygen tensions.

PEEP Level			
5 cm H_2O	**10 cm H_2O**	**15 cm H_2O**	**20 cm H_2O**
78 torr	100 torr	120 torr	65 torr
99 torr	105 torr	101 torr	57 torr
67 torr	87 torr	116 torr	88 torr

A modification of one-way ANOVA allows for homogeneous blocks of subjects to be incorporated in the design, which is termed a *randomized block design.* The design is diagrammed in Figure 7-8. In addition to the *n* levels of the treatment variable, there are blocks, which are groups of subjects who are homogeneous on some trait or traits.

Figure 7-8. Randomized block ANOVA.

	Experimental Factor (independent variable)			
	Level 1	**Level 2**	**Level 3**	**Level n**
Block 1	subjects	subjects	subjects	subjects
Block 2	subjects	subjects	subjects	subjects
Block 3	…	…	…	…
Block n	subjects	subjects	subjects	subjects
	mean 1	**mean 2**	**mean 3**	**mean n**

For example, if the independent variable is a bronchodilator, the dependent variable is forced expiratory volume at one second (FEV_1) and we suspect that age can affect the dependent variable even using all males. The design minimizes the effect of age, the nuisance variable, by blocking on age. Subjects might be grouped into blocks with homogeneous ages (Figure 7-9). For instance, block 1 is 20 to 29 years old, block 2 is 30 to 39 years old, and so forth.

Figure 7-9. Example of a randomized block ANOVA design.

	Bronchodilator Drug		
	Dasage 1	**Dosage 2**	**Dsage 3**
age 20-29	mean FEV_1	mean FEV_1	mean FEV_1
age 30-39	mean FEV_1	mean FEV_1	mean FEV_1
age 40-49	mean FEV_1	mean FEV_1	mean FEV_1

Subjects within a block are then randomly assigned to the *n* treatment groups. The number of subjects within blocks should be equal. Homogeneity may also be achieved in animal studies by use of littermates within a given block. A special case of blocking is to have a single subject in each block, with *repeated measures* of the subject for each treatment level. The order of treatment levels is randomized for each subject, and the subject should be in the same condition when each treatment level is experienced that is, any residual effects from the previous treatment level must have ended before a new treatment level begins. When it is desirable to have the same group of subjects undergo two or more treatment levels, the randomized block used as a repeated measures design can be very useful.

The randomized block design is most appropriate when there is greater variability among blocks than *within* blocks and when this variability makes a difference in the dependent variable. The creation of blocks allows the researcher to isolate the effect of the independent variable and statistically remove the variability in the dependent variable due to the nuisance (blocking) factor.

A *completely randomized factorial design* can be understood as an extension of the previous two designs. This design is diagrammed in Figure 7-10. The term *factorial* experiment indicates that more than one independent variable is being investigated. Figure 7-10 illustrates the design for a 2 x 3 factorial. Each number refers to the number of levels of an independent variable. For example, in Figure 7-10, which shows a 2 x 3 design, there are two levels of factor 1 and 3 levels of factor 3. In a 2 x 4 x 2 scheme, there are three independent variables, with two, four, and two levels, respectively.

Notice that the treatment variables are "crossed," that is, all combinations of treatments can be considered. Subjects are randomly assigned to treatment combinations, or cells, with different subjects in different cells. The advantage of a factorial design lies in the economy of investigating more than one treatment variable in a single study. The design allows the investigator to analyze statistically the effect of each treatment variable, and in particular to detect any *interaction* between the treatment variables.

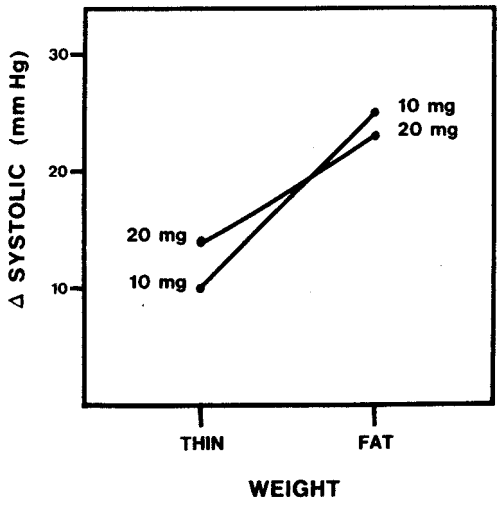
Figure 7-10. A two factor ANOVA design.

Factor 2		Factor 1: Level 1	Factor 1: Level 2
	Level 1	subjects	subjects
	Level 2	subjects	subjects
	Level 3	subjects	subjects

An example is helpful to illustrate both a factorial design, as well as the concept and importance of interaction. Figure 7-11 gives a 2 x 2 factorial design in which the two independent variables are drug dosage and weight. There are two levels of each independent variable. The dependent (measured) variable is change in systolic pressure *(*Δ systolic). First, consider the mean values for 10 mg and 20 mg dosages, calculated as (25+10)/2 = 17.5 and (23+14)/2 = 18.5. There is little difference (17.5 mm Hg compared to 18.5 mm Hg). There is a larger difference between fat and thin subjects; (25+23)/2 = 24 mm Hg compared to (10+14)/2 = 12 mm Hg. If we had looked at the independent variables (drug dosage and weight) in two *separate* studies, we might have concluded that there is no difference in 10 and 20 mg dosages, and there is a difference between fat and thin subjects. By combining the independent variables in *one* study, we find they *affect each other,* that is, there is interaction between drug dosage and body weight.

Figure 7-11. An example of two way ANOVA with interaction. Cells contain mean values for change in systolic blood pressure.

	Drug: 10 mg	Drug: 20 mg
Fat	25 mm Hg	23 mm Hg
Thin	10 mm Hg	14 mm Hg



Interaction can be defined as the situation in which one treatment gives different results under different levels of the other treatment. The graph in Figure 7-11 illustrates this. Results with the two dosage levels

are graphed at each level of body weight. The measured variable, change in systolic pressure, is on the vertical axis. Points on the graph are the factorial cell entries. We see that the 20 mg dose gives a larger change in systolic pressure than the 10 mg dose for thin subjects. But the reverse is true for fat subjects. This is interaction: Levels of drug dosage give different results at different levels of body weight. The numbers in the factorial cells show the same result, but they are less striking than the graph. This result, which is clearly important, would not have been detected in two separate studies.

In a factorial design, we distinguish the effect of each treatment variable (main effects) from the interaction effect. Whether main effects or interaction effects are significant is decided statistically. If interaction *is* present, main effects must be interpreted with qualification. In general, when the graph in Figure 7-11 shows parallel lines, there is no interaction between treatment variables. When the lines are not parallel, whether they cross or not, interaction is present, and statistical testing determines if this interaction is significant

Validity of Research Designs

The primary rationale for a research design, particularly experimental designs, is *control,* to answer correctly the question: Is X related to Y? Experimental validity is concerned with the correctness of conclusions in a research design, and is differentiated into internal and external validity.

Internal validity: The extent to which we are correct in our conclusion concerning the relation of independent and dependent variables.

External validity: The extent to which we are correct in generalizing the conclusions from sample results to the population.

Internal validity includes the specific question of *statistical conclusion validity,* since inappropriate statistical analysis or failure to meet the assumptions of statistical tests is so widespread in the medical literature, as well in other fields.

Threats to internal validity exist and challenge any study. Completely eliminating all potential threats is often impossible (because of practical constraints or the nature of the question). The researcher, as well as the consumer of research, should be aware of the following extraneous factors that can cause incorrect conclusions or inaccurate generalizations.

Threats to Internal Validity

- History and maturation: The passage of time introduces the possibility of events affecting subjects, or changes in the subjects themselves, that may affect the dependent variable. For example, news of a stock market crash might worry some subjects and raise their average blood pressure; or change in seasons may affect the severity of asthma episodes as much as an experimental drug.
- Instrument change: Measurements obtained can be affected by changes in instruments. For example, bias caused by friction building up in a water-seal spirometer can change later measurements.
- Mortality: Any loss of subjects, whether voluntary or through death, can seriously affect research results. For example, if less severe asthmatics lose motivation to continue in a treatment, then the results may be inaccurately poor or even indicate no effect of treatment.

Threats to External Validity

- Population validity: The accessible population is not equivalent to the intended target population. The possibility of this occurs when volunteers are used because volunteers may not typify the entire population. *Volunteerism* leads to a self-selection bias and is an inherent difficulty in survey research.
- Hawthorne effect: The awareness of being in a study can alter a subject's responses 6r behavior, even in medical patients. Psychosomatic effects are quite real, as evidenced by stress ulcer. The effect is named for the particular study that discovered it, which took place in the Hawthorne plant of the Western Electric Company. No matter what treatment was employed, productivity rose! There was a positive response to perceived attention. Subjects not in such a study may not respond similarly, and this limits the ability to generalize beyond the study.
- Experimenter effect: Investigators can also bias results by consciously or unconsciously conveying expectations or motivating subjects to respond to treatment more than to placebo. A double-blind approach helps to control this, since the investigator does not know when a subject receives treatment or placebo. With an investigator causing bias in the results, the conclusion cannot be generalized.

The threats to internal and external validity cited do not provide an exhaustive list. The threats given, however, can easily occur in medical or clinical research. The term *validity* used to denote experimental validity is not the same as measurement validity. The latter refers to an instrument measuring what it is intended to measure.

NON-EXPERIMENTAL STUDY DESIGNS

Non-experimental research designs do not directly manipulate or control an independent variable. The main activity is observation. This design is often necessary when the researcher must use intact groups. Such groups are created or defined on the basis of inherent traits or traits that are acquired in the course of natural events. Such traits are termed attribute variables. Examples include smokers and nonsmokers, second-graders in a school, patients having neuromuscular diseases, all 20 year olds, subjects requiring mechanical ventilatory support, males, and so on. The *experimental* variable that is manipulated is replaced by an *attribute* variable or trait in non-experimental research designs. Comparison or control groups can be identified in some cases, defined by the presence or absence of the trait being investigated.

Non-experimental research designs lack the manipulation and control of experimental designs. Groups remain intact when it is not feasible or ethical to manipulate the trait defining the group. Naturally occurring traits, such as age, sex, or presence of disease, obviously cannot be assigned to subjects. With other traits, such as smoking, random allocation to smoking or not smoking is not considered ethical. Ethical considerations will often require the use of non-experimental design studies in medical research. Are we ethical in withholding a possibly beneficial or superior treatment for the sake of a placebo control? Can we decide who receives ventilatory support and who does not, among patients who meet criteria for mechanical ventilation? Historically, a very well known medical experiment did use an experimental design and randomly assigned placebo versus vaccine: The 1954 experiment to determine effectiveness of the Salk polio vaccine. There was discussion at the time concerning the ethics of giving placebo to children who might contract polio. However, part of the research *did* use a randomized

placebo control method, and as a result the researchers were able to conclude that the Salk vaccine was effective. By contrast, using non-experimental methods, it has been impossible to determine conclusively that smoking causes lung cancer.

There is no well established and widely agreed upon classification scheme for the different types of non-experimental research. In addition, terms associated with non-experimental research, for example, *ex post facto, retrospective, prospective,* and *cohort,* are also used in different ways among individual writers, and in the various fields of research such as the behavioral sciences (education, psychology, sociology) or clinical medicine (epidemiological research). With this realization in mind, I offer a description of the major types of non-experimental research, using labels that are acceptable among medical research methodologists. The types of non-experimental research to be presented are the *prospective study*, the *retrospective study*, the *case study*, the *survey*, and the *correlational study*. Non-experimental designs can be very useful and often the best possibility for medically related research, especially in allied health fields. The following types of non-experimental research will suggest that clinical practice can be a fertile source of research information to begin putting health care on a scientific basis.

With non-experimental research, control can be a major hindrance to drawing conclusions. One solution is to match intact groups on certain variables, or use analysis of covariance to statistically equate groups on given variables. Another approach is to incorporate nuisance variables into an analysis of variance design to examine or control their influence.

Retrospective Studies

Retrospective studies attempt to reason from a present effect, or consequence in a population, back to antecedent causes. For example, a retrospective study on lung cancer would identify cases of lung cancer, identify a comparable group based on age, sex, or other characteristics, and then determine what variables were prevalent in the lung cancer group that are not in the control group. Another example is a retrospective study to identify antecedents of breast cancer by profiling and taking extensive histories on subjects with breast cancer. The idea is to find a common characteristic in those subjects with the disease. That characteristic is then interpreted as a risk factor. Causality is almost impossible to establish in such studies; instead, we identify functional relationships between variables

Prospective Studies

Prospective studies attempt to reason from a present antecedent or event in a population to future consequences or effects. A good example is a study on complications of assisted ventilation. Suppose records were kept of complications (mainstem intubation, pneumothorax, and so on) in a group of patients requiring assisted ventilation. The data from such a study could be used in quantifying relative frequencies of common complications.

The term *cohort* has also been used to describe a group of subjects who are followed prospectively, or forward, in time. The defining feature of cohort research is that the group is followed forward from antecedent to outcome. Terms such as *development* or *follow-up* have also been used to describe studies in which an intact group or cohort is followed longitudinally over time.

The terms prospective and retrospective are properly used in reference to the chronological direction in which a group is followed. *Prospective* describes a study where the group of subjects is tracked in the forward direction. *Retrospective* applies when a group is followed in a backward (retro) direction. These terms have been misapplied to describe the direction in which the data are collected. For example, a

researcher wishes to study the complications of mechanical ventilation in a homogeneous group of chronic obstructive pulmonary disease (COPD) patients. Suppose that all such patients who require mechanical ventilation next year are identified for inclusion. If the researcher plans now to follow such a group, then observations can be made from the time of commitment until discontinuance of the ventilator occurs. On the other hand, if the researcher decides now to investigate the same group of patients from 5 years ago, then subjects as well as complications will have to be gleaned from medical records in the past. Regardless of whether we start collecting data now or review old records, we are following the group in a forward direction, from ventilator commitment to discontinuance. Such a study is prospective, whether the data are obtained at the time of ventilatory support or afterwards. In other words, it is not the chronology of data collection but the direction of population pursuit (forward versus backward) that distinguishes prospective from retrospective studies.

Case Studies

The case study usually provides a description of a single subject, with a treatment or trait (e.g., disease state) that is unusual, rare, atypical, or experimental. Case studies often provide a good teaching example.

Surveys

A survey typically gathers information on a large group of subjects or units, by either written or oral questionnaires, or even on-site inspection. A sample survey based on random methods can give very good information on populations, as evidenced by political polls and quality control techniques that use random checking. A manpower survey in a particular state such as Georgia can indicate what proportions of health care practitioners are registered, certified, or on-the-job trained. Problems with survey questionnaires include response rate and self-selection bias (volunteerism). Unambiguous wording, clear-cut and non-overlapping answer choices, a brief, clear introduction that motivates the reader to respond, and an uncomplicated format help to obtain accurate and adequate responses. Pilot testing of a questionnaire to eliminate ambiguities is essential. This is done by administering a small number of questionnaires to volunteers who can provide some feedback on the clarity and appropriateness of the questions.

Correlational Studies

For a technologically oriented field, correlational studies are very useful and provide information on the presence and strength of a relation between two variables. For example, we might ask if there is a correlation between mid-maximal expiratory flows and PaO_2. We would select a group of subjects who are homogeneous in representing a population, and then in each subject measure the mid-maximal flow and the PaO_2. Then the pairs of measurements from each subject are statistically analyzed to determine the strength or absence of correlation between the two variables (flow and PaO_2).

The term *correlation* means that two variables co-vary. Co-variance means that when one variable changes, the other variable is also likely to change (in the same or opposite direction). *Correlation does not imply causality*. Both variables may be caused by a third variable, and they will co-vary simultaneously.

Actually, any two variables can be analyzed for the presence of correlation. For example, a suspected relationship between degree of sunspot activity and brightness of the northern lights might be analyzed using a correlation study. Generally, theory should suggest the possibility of a relationship between two variables. This occurs usually when there is a cause-effect relation, or if two variables already coexist in

the same subject. Examples of the latter include situations where the same entity is measured by two different instruments and one uses a correlation study to examine the accuracy of one instrument against the other; or where physiological variables such as heart rate and blood pressure exist in the same subject.

QUESTIONS

Definitions

- Assessable population
- Target population
- Sample
- Variable
- Independent variable
- Dependent variable
- Nuisance or confounding variable
- Placebo
- Hawthorne effect

True or False

1. Random sampling is essential for descriptive statistics but not for inferential statistics used in hypothesis testing.
2. In simple random sampling, every unit in the population has an equal chance of being selected.
3. The *Law of the Single Factor* states that if two situations are equal except for one factor, any difference between the situations is attributable to that factor.
4. Pre-experimental research designs are preferred because there is full control of nuisance variables.
5. The pre-test/post-test design is classified as pre-experimental.
6. The case control design is considered quasi-experimental.
7. The randomized control design is the only true experimental design.
8. One advantage of ANOVA is that more than two groups can be compared at a time.

Multiple Choice

1. Which of the following non-experimental designs attempts to reason from present effects back to antecedent causes?
    a. Retrospective study

b. Prospective study
c. Case study
d. Survey
e. Correlational study

2. Which type of study seeks information on the strength of a relation between two variables?
   a. Retrospective study
   b. Prospective study
   c. Case study
   d. Survey
   e. Correlational study
3. Which type of study attempts to reason from a present event to future effects?
   a. Retrospective study
   b. Prospective study
   c. Case study
   d. Survey
   e. Correlational study
4. Which type of study usually provides a description of a single subject that is unusual and provides a good teaching example?
   a. Retrospective study
   b. Prospective study
   c. Case study
   d. Survey
   e. Correlational study
5. Which type of study gathers information from groups of subjects using questionnaires?
   a. Retrospective study
   b. Prospective study
   c. Case study
   d. Survey
   e. Correlational study

SECTION III CONDUCTING THE STUDY

Chapter 8. Steps to Implementation

Although the details of formulating a research plan have been discussed earlier, it cannot be overemphasized that careful, detailed planning should be completed long before the study begins. Devising a comprehensive plan should be the first step of any research project. Havin a plan will avoid many problems during the actual study.

WRITING THE STUDY PROTOCOL

Three major benefits accrue from a *written* research plan. First, the process of writing it out will help you clarify the goals of the study and methods of investigation. The realization that problems in approach or analysis exist may not become clear until ideas are committed to paper. Second, you must often present a plan to obtain permission or approval to proceed with the study. Permission may need to be sought from a funding source, institutional review board, department manager, or student advisor before a study may begin. Third, the research plan, or *protocol*, as it is often called, provides an operational guide for the entire research team. Successful coordination of study personnel is all but impossible without a detailed protocol. For these reasons, a properly formulated proposal is an essential first step in the research process.

Creating a General Plan

Organizing the plan takes a good deal of thought. To place the plan in proper perspective, you should be able to answer the following questions:

- What is the primary goal of the study?
- What makes this particular study important?
- What has previous research shown in this area?
- How is the study to be carried out?

You need concise answers to these questions to help clarify the objective and methods. The research protocol is a map of the path the project is to follow. It must be written in language comprehensible to all team members and should clearly delineate the role of each participant. It should make clear that the people involved with the study know their responsibilities, that the study will produce valid results, and that an interpretation will flow easily from results obtained.

A well-written protocol must address several issues. For example, it should place special emphasis on the criteria for selection of the experimental population sample. It should also clearly specify which factors will be used to classify subjects into groups and must define the criteria for treatment. Here is a general outline you can follow:

1. List the study's specific aims. These should include a statement of both what goals are to be accomplished and the hypothesis to be tested. You must condense the broad research topic into a concise problem statement.

2. State clearly the primary significance of the study. By this time you should have completed your literature review. Evaluate your project in light of the published data. Show how your study will fill in knowledge gaps. Relate the specific study objectives to long-term goals. For example, your study objective might be to assess the accuracy of oxygen analyzers when using concentrations below 21%. The long-term goal of such a study would be to ensure the safety of providing sub-atmospheric oxygen levels to children with hypoplastic left heart syndrome.

3. Cite any preliminary studies you may have undertaken. Show that preliminary experiments have demonstrated the feasibility of methods you intend to use for the current study. This would also be the place to state the qualifications (as needed) of all your co-investigators. For example, if you are studying the contamination rate of nebulizers in the home, you might want to include a microbiologist.

4. Specifically state the experimental design. Several points to consider:

a. innovative procedures

b. advantages and disadvantages of methods used

c. limitations of the methods or study design

d. anticipated difficulties

e. alternative approaches in case of experimental obstacles

f. sequence of experimental events (make a flow chart if necessary)

g. procedures used to analyze the data

Don't be discouraged if several revisions are required to get the plan right. And always consult your mentor.

The IRB Study Protocol Outline

If you are planning a study involving human subjects, you will have to create the study protocol using a detailed format, specified by your hospital's Institutional Review Board (IRB). An example of the format is as follows:

1. Name of Investigator/Co-Investigator(s)

The principal investigator must be a member of the hospital staff in all studies that involve therapeutic interventions or that alter medical care.

2. Title of Project

All projects must have a title without abbreviations. This title must appear on all pages of the consent form.

3. Introduction

This section should include the relationship of the research to previous studies in the field (including pertinent references) and the significance of the study. Relevant laboratory and/or animal studies should be mentioned.

4. Purpose, Specific Aims and Hypotheses

This section should state clearly what is hoped to be learned from the research.

5. Study Design

This section is to inform the IRB of the specific nature of the procedures to be carried out on human subjects in sufficient detail to permit evaluation of the risks. This section should also provide information that will allow the IRB to confirm the claim that methods employed will enable the investigator to evaluate the hypothesis posed and to collect valid data. The study design and specific procedures must contain the specific information needed in the consent form and to allow the evaluation of the consent form.

The section on study design should include a brief discussion of the following:

a. Specific Procedures

All specific procedures to be performed on human subjects for purposes of research should be detailed. Uncommon medical procedures should be fully explained. Distinguishing between the usual patient care and any experimental procedures is important. The protocol should indicate what changes in medical care will occur as a result of the study, how care will be managed, and by whom. A tentative time schedule for various procedures should be provided showing what a subject might expect regarding how long each aspect of the study will take, the frequency and timing of ancillary procedures (i.e., NPO status) and the duration of discomfort. Present complicated studies using a simple flow chart to enhance the narrative description. The location of the study, including laboratories for special testing, must be indicated. In studies involving the use of a placebo or "wash out" period, the protocol and the consent must discuss what will happen if the subject's condition deteriorates.

b. Population

The source and means of identifying patients/subjects and control subjects should be indicated, as well as the number of subjects to be studied. If the study involves patients whose care is the responsibility of departments or special care areas other than that of the responsible investigator, that department or special care area should be identified as having approved the protocol. Protocols must be precise as well as concise in defining a study population and the mechanism whereby the population will be contacted. If contact is being solicited through paid advertising, the ad copy must be approved by the IRB. Specific justification for the use of subject groups with compromised ability to give consent should be described. Such groups are:

1) Prisoners, for whom conditions of confinement, penury and parole may constitute significant duress.

2) Minors: Informed consent is required from a responsible parent or guardian for minors ages 17 or under. In older children, particularly those 13 and older, the investigator is urged, in addition, to obtain assent of the minor on the consent form. The investigator may wish to use a separate "consent form" for minors able to follow it. Consents need to be written very carefully if a minor is to also sign. Any child capable of signing his name should do so. In young children, it may be desirable for one of the parents to be present during an investigational procedure.

3) Legally incompetent people: Studies involving legally incompetent individuals will require the signature of a legal guardian and not the next-of-kin. When competency is in question, the responsible physician will assist in assessing the patient's competency. *Signature of spouse or next-of-kin does not constitute agreement to participate in research studies unless they are legal guardians.*

4) Unconscious patients may not be used in research studies unless there is no alternative method of approved or generally accepted therapy that provides an equal or greater benefit. The investigator and an independent physician must make the determination in writing for the medical record and sign the consent form, if:

   a) The study offers potential benefit to the patient.

   b) The patient's condition is life threatening.

   c) There is inability to communicate with the patient.

   d) There is insufficient time to obtain consent from a legal representative or there is no legal representative.

   e) The next of kin are informed, and agree to the study and sign the consent. However, this does not constitute informed consent. When able the subject has the right to withdraw consent.

5) House staff, students and employees, when directly solicited to participate in a study, must be informed that their participation in a study or refusal to do so will in no way influence their grades or subsequent recommendations.

   Medical students on a clerkship should not be asked to take part in experimental procedures conducted by that service while the student is on that service.

   Employees of the hospital must never be made to feel that their job, promotion, salary, or status in any way depends on participation in research studies.

c. *Financial Considerations*

1) Compensation to subjects.

Experimental subjects may be reasonably reimbursed for their time, their expenses, and the degree of discomfort they may experience during an investigation. Amounts must be specifically stated in both the protocol and consent form and justified in the protocol. Payments should never be so large as to induce a subject to submit to experimentation that they might otherwise reject. Payments totaling $600 (subject to IRB regulations) or more per year will result in a 1099 tax form being sent to the subject. Compensation should never be considered as a benefit of the study.

2) Extra Costs Incurred for Purposes of the Study

The IRB is concerned about the cost accounting of research studies, and investigators must address the problem of extra costs incurred because of the research project. If costs due to research are to be incurred by the patient/subject, such costs must be stated on the consent form. It is illegal to charge non-therapeutic studies to the patient or third party payers.

Examples of procedures in this category are:

a) cost of radiographic studies done solely for research purposes

b) cost of additional anesthesia time in a surgical procedure which includes research procedures

c) cost of techniques/drugs where there is no benefit

d) the FDA classifies investigational devices as to whether or not they can be charged for. Many devices are chargeable.

e) drugs being used under an IND number may not be charged to patients or third party payers.

The protocol and consent form must include a statement regarding the responsibility for costs.

d. Risks and Benefits

1) The risks and potential benefits, if any, of the proposed study to the patient/subject, his family, and/or society, should be discussed, as well as an analysis of the risk/benefit ratio.

2) A discussion of precautions to minimize risks is appropriate so that the IRB can ascertain the true risks of procedures to be performed. The precautions range from those applicable to a group, such as the exclusion of pregnant or potentially pregnant women from a study, to those applicable to an individual subject, such as the presence of an emergency cart in those studies in which a patient may be subject to arrhythmias.

3) The investigator should offer justification for the proposed study. How significant is the new knowledge being sought in relation to the potential risks in carrying out the research?

4) Studies involving children. The approval of all studies involving minors requires the identification of which OHRP risk categories apply to that protocol.

6. Consent Form

The consent form should express the realistic expectations of participation in the research study, avoiding inducement by raising false hopes. The consent form should be written in simple, non-technical lay terms. Medical jargon should be avoided. The consent form should be written in a manner that is understandable to an individual who has completed an eighth grade education. The consent form is to be written in first person, i.e., I understand that I am being asked to voluntarily participate in a study . . . In consent forms for studies involving minors, statements should read "I/my child" when appropriate and modified depending on the age of the child and the child's ability to understand. Consent forms should not include listings of inclusion or exclusion criteria.

The following should be included:

a. Purpose of the Study and Individual Participation

A clear and concise statement of the purpose of the research study and why the individual is being asked to voluntarily participate in the study should be indicated. Participation of normal individuals as control subjects must be identified. For patients who have an identifiable responsible physician, a statement in the consent form that his or her physician has given approval to contact this patient for possible participation in a study is mandatory.

b. Study and Procedures

The study and procedures to be performed should be described. Identify or separate routine management or therapeutic procedures from procedures being carried out solely for investigational purposes. If a placebo is to be used in a study, identify this and define a placebo early in the description of a study or in stating the purpose of a study. A statement indicating reasonable expectations of time commitment must be included. If blood is being collected as part

of the study, the amount of blood should be indicated in lay terms (teaspoons/ tablespoons) as well as risks involved, and what will be done with the blood. *The lack of a clear statement separating routine care from research has been the most common reason requiring revisions in consent forms.*

c. Risks and Benefits

Risks and discomforts of each procedure should be included, as well as degree of risk. Drug side effects should be stated and whether they are rare or common. Where appropriate, a statement indicating possible unforeseen side effects should be added. Benefits should be stated clearly. Distinctions between personal and societal benefits should be indicated. The consent must address the worsening of a condition or lack of response in treatment protocols or in protocols involving a medical intervention. This would include criteria for stopping a study particularly if a study involves a placebo or a "wash out" period. In many protocols that state "there may be a benefit" it would be more appropriate to state "there may or may not be a benefit". Where appropriate, there should be a statement regarding compensation and medical care in the event of research injury.

d. Alternatives and Withdrawal

Alternatives, including non-participation, must be stated when a new diagnostic or therapeutic procedure is being used. Discussion of the alternatives must be fair and should balance the alternatives against the proposed experimental therapy or procedures. The patient must be informed that he/she may withdraw from the study at any time.

e. Treatment After the Study

Discuss treatment or management after completion of the study in studies that involve therapeutic interventions.

f. Financial Considerations (Cost Responsibility Statement)

When appropriate, the consent statement must indicate compensation to subjects. Extra costs that may be incurred as part of participation in the study should be spelled out. Otherwise, indicate that no extra costs are involved. A statement in the consent form should acknowledge that third party payers may or may not cover the expense of some procedures and hospitalization involving research studies. It is mandatory that cost responsibility for usual routine care be separated from cost of research procedures. When appropriate, a statement may be added to the "cost statement" in the consent form, "I should check with my insurance carrier to determine what my coverage will be".

7. Confidentiality Statement

This statement is presented in the final paragraph of the consent.

8. Identification of Persons Obtaining Consent

The persons obtaining consent must be identified, including a phone number. Include home telephone numbers. In all studies involving therapeutic interventions or significant risk, the principal investigator, physician co-investigator or a specific research nurse shall obtain consent. Technicians or "floor nurses" may not obtain consent. NOTE: The principal investigator is required to co-sign the consent document within two weeks from the time consent is obtained from the subject.

As you can see, a study involving human subjects can be very complex. That is why most novice researchers start with device evaluation studies that can be performed outside of patient care areas.

FUNDING

Funding should never be an obstacle for the novice researcher beginning with a small project. Most studies produced in health care departments require no special funds, with the possible exception of some overtime for the staff involved. If the study is a device evaluation for something new on the market, the manufacturer will usually be glad to give you free samples to test, assuming your protocol is convincing. After all, when you publish your results the manufacturer will get free advertisement.

Major clinical studies are generally funded by grants from the government, but health insurance agencies, nongovernmental public institutions, the pharmaceutical industry, and the medical equipment industry also support clinical studies. The educational credentials required to obtain research grants may be quite demanding. Successful researchers, through their protocols and grant applications, must demonstrate the following abilities to:

- think clearly and logically
- express logical thought concisely
- discriminate between the significant ant the inconsequential
- display technical prowess
- handle abstract thought
- analyze data objectively and accurately
- interpret results confidently and conservatively.

American Respiratory Care Foundation

One source of funding for respiratory therapists is the American Respiratory Care Foundation. The American Respiratory Care Foundation is dedicated to furthering the art, science, quality and technology of respiratory care. The Foundatin is a not for profit organization formed for supporting research, education, and charitable activities. Specific activities of the Foundation include funding clinical and economic research, education, recognition awards, educational activities, literary awards, and scholarly publications. The Foundation is deeply committed to health promotion, disease prevention, and improving the quality of our environment. It seeks to educate the public about respiratory health and assists in the training and continuing education of health care providers.

The ARCF supports a number of grants available to respiratory therapists. More information can be found on their website at http://www.aarc.org/arcf/main.html.

DATA COLLECTION

One of the most important considerations for clinical studies is the method used for data collection. There must be a plan for how and by whom the data will be recorded. Data collection is the most time consuming and expensive step in the research process. If the data are misplaced or cannot be understood when it is time to analyze the study results, the project is ruined.

The Laboratory Notebook

Good scientists keep a notebook in which they record the permanent written record of all activities related to the research project. The act of writing in the notebook causes the scientist to stop and think about what is being done, encouraging "good science". And you never know, you may discover something useful that you want to patent. The patent system in the United States rewards the 'first person' who invents a new product. The laboratory notebook helps you prove you were first.

The information in the notebook is used for several purposes. Most importantly, it records the experimental data and observations that will be used later to make conclusions. Everything you do and the sequence in which you did it should be recorded, because you often can't determine what will be important later. Include drawings of experimental setups and flow diagrams of the sequence of events. Include tables of measurements descriptions of procedures. Problems and limitations encountered are just as important to record as the successful experimental outcomes. See Table 8-1 for more ideas.

Table 8-1. Ideas for keeping a laboratory notebook.

- Detailed records of the concepts, test results and other information related to the experiment should be kept. You can start from the very first moment you think of an idea.
- Ideas, calculations and experimental results should be entered into the notebook as soon as possible, preferably the same date they occur
- All entries should be made in the notebook in permanent black ink and should be as legible and complete as possible. Do not use abbreviations, code names or product codes without defining them clearly.
- Draw a line through all errors. *Do not erase*.
- Entries should always be made in the notebook without skipping pages or leaving empty spaces at the bottom of a page. If you wish to start an entry on a new page, draw a line through any unused portion of the previous page. Never tear out or remove a page from the notebook.
- You can buy a specially printed laboratory notebook or make one yourself. Use a bound notebook, because the pages cannot be added or subtracted without that being evident.
- Number all your pages consecutively. On any blank pages or portion of a page left, you should draw a line across. Start a new notebook when yours is full. Each notebook should be assigned a consecutive number.
- Keep your notebooks in a secure location and make records of when you take or return your notebooks from that spot.
- Use a header for each entry with the following information - date, project number, subject, participant(s).
- The more details the better, make sure that you have all the information you will need when writing up the results later.
- Make records of everything. Include all your tests, not just the successful ones. Add all your sketches, measurements, and computations.

Table 8-1 continued. Ideas for keeping a laboratory notebook.

- All loose material, such as drawings, data collection forms, printouts, photographs, etc., should be signed, dated and cross referenced to a particular notebook entry.
- If you can, tape or staple the loose material into the body of the appropriate notebook entry.
- Anything else such as samples, models, prototypes, etc., should be carefully labeled with a date and cross referenced to notebook entries. Keep all of it.

The guiding principle for note keeping is to write with enough detail and clarity that some scientist in the future could pick up the notebook at some time, repeat the work based on the descriptions, and make the same observations. If fact, that scientist in the future might just be you after you have forgotten what you did!

A professional laboratory notebook is bound with specially printed pages (Figure 8-1). You can purchase notebooks at university bookstores or through the mail. You should be able to find many sources on the Internet (see for example www.laboratorynotebooks.com/). Look for features like sequentially pre-printed numbered pages, spaces for you to sign and date, and instructions on how to use the journal to record your observations. Also look for pages with blue-lined grids for easy drawing. Some notebooks have special copy features; copy drawings on a light copier setting and the grid pattern fades away for preparing manuscript drawings. Never use a loose leaf notebook. Never buy 3-ring binders to use as a notebook. Never buy a legal pad or any glued together notebook. Buy a notebook with pages as secure as possible - a bound or sewn notebook. Otherwise, pages will get accidentally torn out or out of order. Mead brand composition books are acceptable if you can't find a true laboratory notebook. Buy only notebooks with white pages - the lines can be colored blue or black. All entries are made in ink with errors crossed out rather than erased. Figure 8-2 is an example of the instruction page from an industrial laboratory notebook.

Figure 8-1. Example of a page from a laboratory notebook.

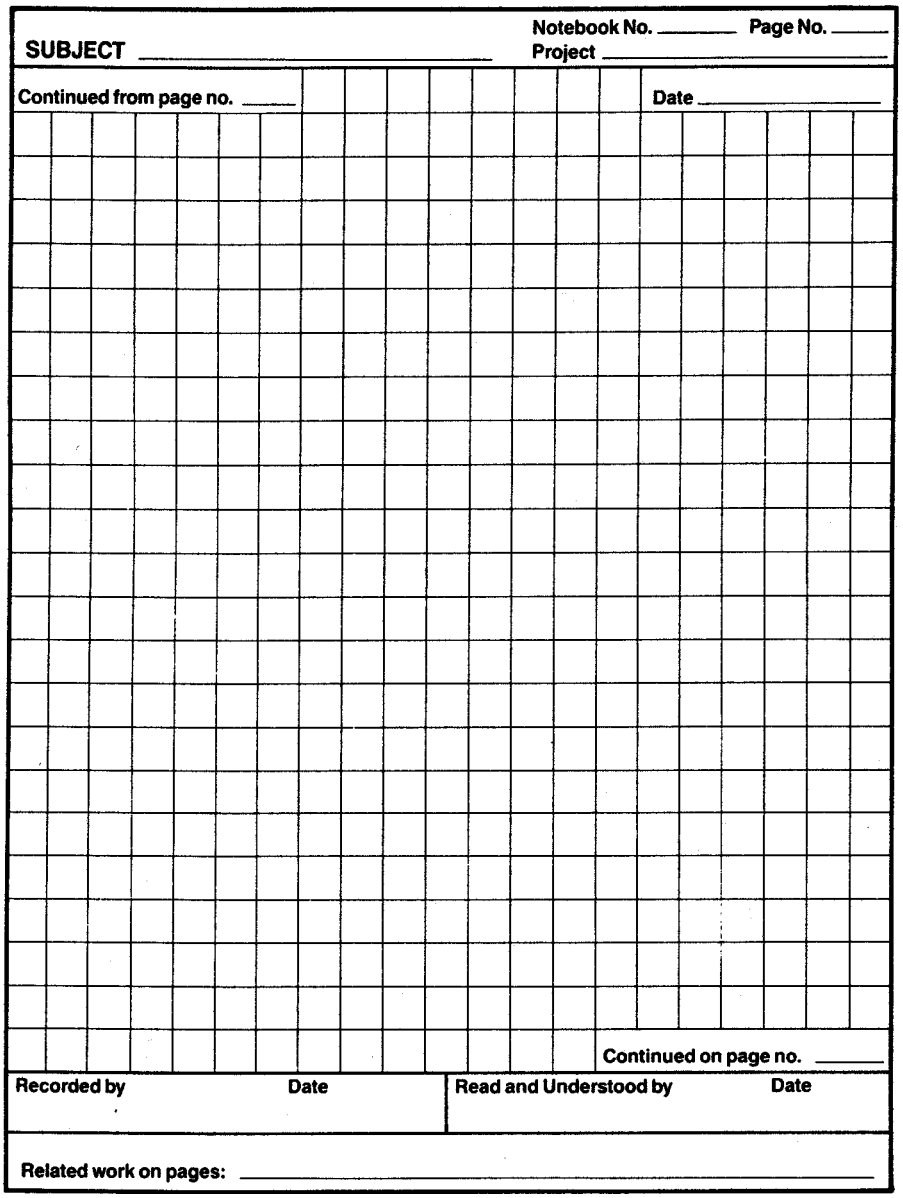


Figure 8-2. Example instructions from an industrial laboratory notebook.

LABORATORY NOTEBOOK

INSTRUCTIONS

This Laboratory Notebook is the property of [The Company]. It is assigned to you so that you may keep a complete, careful, chronological record of your work. The work which you do and the data which you enter in this book are confidential; they must not be disclosed to unauthorized persons. The Notebook must not be removed from the laboratory premises. Its preservation and maintenance are your responsibility; in case of damage, loss or disappearance, report the facts to your section manager at once.

The purpose of each entry in your Laboratory Notebook is to provide a complete record of your work, one that would enable one of your coworkers to repeat, if necessary, exactly what you did and secure exactly the same results, without having to ask any questions of anyone. You will find these specific instructions helpful in preparing entries which will meet this requirement.

1. Plan your experiment carefully, and plan the presentation which will best record the data you expect to secure. Since the duplicate page will be extracted from the Notebook and attached to the appropriate progress report for filing in the project file, data on each page must be limited to one specific project.
2. After Title and Project number, date and the objective have been filled in, your entry should record (1) the purpose of the experiment; (2) the materials used and their quantities; (3) the apparatus; (4) the procedure and manipulation (times, temperatures, pressures, pH's, and the like); and (5) the results. Where procedure or apparatus is standard, it is sufficient to describe it by reference; for example, ASTM D236-54T, or by reference to an earlier notebook page where it was fully described.
3. All data is to be recorded directly into the Notebook. Recording of original data on loose pieces of paper for later transcription into the Notebook is to be avoided. Should use of loose paper be necessary for proper conduct of an experiment, the loose paper should be signed and dated, and cemented into the Notebook. The data must be transcribed into the Notebook on the same day it was taken, and the Notebook entry should refer to and identify the loose paper which has been cemented into the book.
4. All entries must be made in ink. Erasures are not permitted. If a mistake is made, draw a line through the erroneous material and make a corrected entry immediately following.
5. Every entry must be dated, and signed at the foot and at the end of each day. In no event may the entries be signed less frequently than each page.
6. If, for clarity of presentation, it is desired to start the new entry on a new page when the previous page has not been entirely filled, draw a diagonal line across the unused portion of the page.
7. Pages must be used consecutively. Leaving a page or pages blank for later use is absolutely forbidden. Entries must be presented in chronological sequence.
8. If one of your coworkers (not a codiscoverer of the subject matter) has witnessed an experiment you have conducted, to an extent that enables him to state of his own knowledge what you did and what results you secured, have him sign and date the Notebook record of the experiment under the legend "Witnessed and understood by". If the experiment seems to you to be of sufficient importance, arrange to have it witnessed.
9. Pages are provided in the front of this book for an index to the subject matter covered. The index pages should be completed as the work progresses to afford ready access to the data recorded.
10. Avoid stating conclusions, particularly of failure or abandonment. Let the results speak for themselves.
11. When this book is filled, or upon termination of your employment, it must be turned in to your section manager.

Book No. _____ Assigned to _____ Date _____

Returned to Section Mgr.—Date _____

Specialized Data Collection Forms

Sometimes data collection is easier on specially made forms instead of in the laboratory notebook. This is certainly true if other people in other locations will be collecting the data. The best way to make forms is to use a spreadsheet program like Microsoft Excel to design a form you can print on paper. You can make the form look any way you want *and* build in equations to summarize the data (e.g., calculate sums, means, and standard deviations). You can even do graphing and statistical analysis within Excel, although dedicated statistical software is easier to use. But for that, you can either export the data from Excel or simply "cut and paste" the data from Excel into the statistics program.

Computers

You can hardly be a scientist in today's world without some computer skills. At minimum, you need to know how to type, how to use word-processing and spreadsheet software and preferably, how to use statistical and database software. You need to know the basics of how to store data in files and how to make backup files. Get some experience with the Internet and have an email address. The Internet provides a way to access a whole world of information, not only for literature searches but to get help with every aspect of research. In addition, many journals are now accepting electronic submissions of abstracts and papers. And don't forget that help from your colleagues or mentors is just an email away.

One area of computing that merits special attention is the use of personal digital assistants (PDAs). These little handheld computers can be used to collect data in the laboratory or on patient care divisions much more conveniently than paper forms. They are designed to share or "upload" their data with your desktop computer where you would normally perform the data analysis. PDAs like the Palm computer (www.palm.com) can be programmed with forms, spreadsheets, and even databases with minimal effort. The can even be connected to specialized sensing devices to automatically record signals from experiments. These little gadgets can really make your life easier.

QUESTIONS

True or False

1. One of the major reasons for writing a study protocol is that it is required to obtain permission from the IRB.
2. Most IRBs will allow the use of any type of outline for a protocol so long as it includes both methods and a risk/benefit analysis.
3. Funding should never be an obstacle for the novice researcher beginning with a small project.
4. A laboratory notebook is no longer necessary now that computers can perform statistical analyses.
5. An IRB protocol will require a statement about any financial compensation to study subjects.
6. A description of risks and benefits can be omitted if the study subjects are prisoners.
7. The consent form should be written in a technical style so the subject's referring physician can interpret the feasibility of the study.
8. You should record everything you do and the sequence of events in a laboratory notebook because you can't always tell what will be important later.

Chapter 9. Making Measurements

Measurements are made either by *direct comparison* with a standard or by *indirect comparison* using a calibrated system. Measurements of length and weight are examples of the direct comparison of an object with an accepted standard (e.g., a ruler or standard mass). The monitors in an ICU typically employ indirect comparison. They convert some physical quantity, like pressure, to an intermediate variable, like voltage, through a relation previously established by comparison to a standard. Ideally, the standard should be traceable (through three or four generations of calibration copies) to the prototype kept by the National Institute of Standards and Technology (formerly the National Bureau of Standards).

BASIC MEASUREMENT THEORY

Every measurement is assumed to have errors. Even standards are simply the best estimate of a true value made from many carefully controlled measurements, Errors fall into two categories: systematic and random.

Systematic errors occur in a predictable manner and cause measurements to consistently under- or overestimate the true value. They can be constant over the range of input values, proportional to the input value, or both. Systematic errors are not affected by repeated measurements but can be reduced by proper calibration.

Random errors occur in an unpredictable manner due to uncontrollable factors. They cause measurements to both over and underestimate the true value. As the number of repeated measurements of the same quantity increases, random errors tend to sum to zero. Random errors often exhibit a normal or Gaussian distribution. This assumption, along with the Central Limit Theorem of statistics provides the basis for establishing the probability of a given measurement value and hence our confidence in the reliability of our observations.

The effects of measurement errors may be expressed as

$$measured\ value = true\ value + (systematic\ error + random\ error)$$

The observed measurement is seen as the sum of the true value and the measurement errors. The goal is to identify and minimize the measurement errors. Calibration does not improve random error.

Accuracy

Accuracy is usually defined as the maximum difference between a measured value and the true value (what we have called error above), and is often expressed a percentage of the true value:

$$accuracy(\%) = \frac{measured\ value - true\ value}{true\ value} \times 100$$

Some authors talk about accuracy as reflecting only systematic error. They define accuracy as the difference between the true value and the mean value of a large number of repeated measurements (which is the definition of *bias* in statistics). Equipment manufacturers generally include systematic and random errors in their "accuracy" specifications as the worst-case estimate for a given reading.

Accuracy is commonly expressed as a percentage of the full-scale reading (Figure 9-1A), indicating a constant error. For example, suppose a device with a scale of 0 to 100 has a stated accuracy of plus or minus 2 percent (written as ± 2%). Two percent of 100 is 2. This means that if the device is used to measure a known true value of 80, the expected error would be ± 2 so the instrument's reading would be somewhere between 78 and 82. Alternatively, the accuracy specification might be stated as a percentage of the reading. (Figure 9-1B), indicating a proportional error. In this case, 2% of the known value is 0.02 x 80 = 1.6, so the reading would lie somewhere between 78.4 and 81.6. Sometimes the accuracy specification includes both full scale and proportional components (Figure 9-1 C). If the accuracy specification does not state which type it is, we usually assume it to be a percentage of full scale.

Unfortunately, the commons usage of the term accuracy is counterintuitive. An instrument that is considered highly accurate will have a low value for its accuracy rating and vice versa. The terms *inaccuracy* or *total error* are more descriptive. Manufacturers do not use these terms, however, because they are afraid it will make their products seem defective.

Error specifications indicate how far the instrument's reading is expected to be from the true value. Inferring the true value from the instrument's reading is not the same problem. For a more detailed discussion, see the section entitled Interpreting Manufacturers' Error Specifications in Chapter 10.

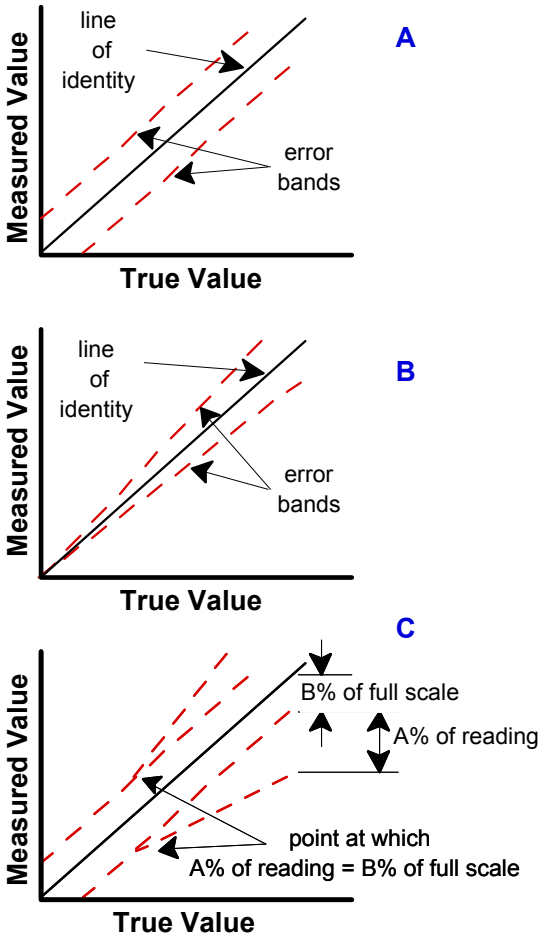


Figure 9-1. Various conventions used to express instrument inaccuracy specifications. The line of identity represents a perfect match between measured and true values, or zero error. The error bands show the range of measured values expected above or below each true value. In other words, the vertical distance from the error band to the line of identity is the expected measurement error.

A. Error expressed as plus or minus x percent of full scale. **B**. Error expressed as plus or minus x percent of the reading. **C**. Error expressed as percent of full scale or percent of reading, whichever is greater.

Precision

Repeated measurements of the same quantity result in small differences among the observed values because of random error. *Precision* is defined as the degree of consistency among repeated results. It is quantified with statistical indexes such as variance, standard deviation, and confidence interval (described in the section on basic statistical concepts). As with the term accuracy, the common usage of the term precision is counterintuitive. A measurement considered to be highly precise has a small deviation from the true value and vice versa

Precision should not be confused with *resolution*, defined as the smallest incremental quantity that can be measured. Resolution is an inherent but often overlooked limitation of digital displays. A digital display changes only when the measured value varies by some minimum amount. Any variation less than this threshold is ignored. For example, digital pressure monitors on ventilators display increments of 1.0 cm H_2O. When used to make repeated measurements of, for example, the baseline airway pressure level, they may give very precise (i.e., unvarying) readings. But they do not have the resolution to detect the small changes in pressure (less than 1.0 cm H_2O) caused by small inspiratory efforts or vibrating condensation in the patient circuit (see *Range Error* under the section on Sources of Bias). If these phenomena were of interest, such a measuring device would be inaccurate.

Inaccuracy, Bias and Imprecision

To avoid any confusion regarding nomenclature, the term *inaccuracy* is used hereafter to mean the total error of a measurement, *bias* to mean systematic error, and *imprecision* to mean random error. Therefore, a highly inaccurate measurement is one that is highly biased and/or highly imprecise. Thus,

$$\text{total error} = \text{measured value} + \text{true value}$$
$$= \text{bias} + \text{imprecision}$$

If the inaccuracy for a given measurement is a positive number, we say that the measured value overestimates the true value and vice versa. We interpret an inaccuracy specification as meaning that any measurement of a known value will be within the given range with a given probability. The effects of bias and imprecision on measurements are illustrated in Figure 9-2.

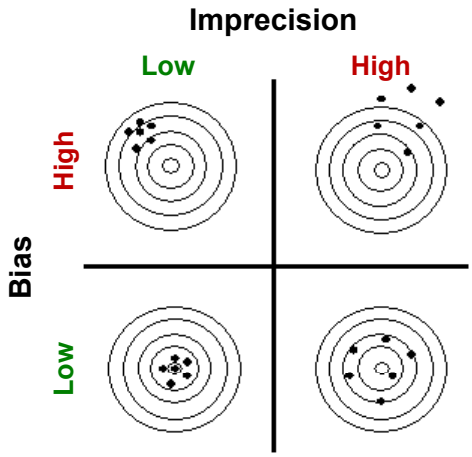


Figure 9-2. An illustration of the effects of bias and imprecision (systematic and random errors) using the analogy of target practice on a rifle range. When bias is low, measurements group around the true value (represented here as bullet holes clustered around the bull's eye on bottom two targets). When imprecision is low, the cluster is tight, showing that the random errors of repeated measurements (or rifle shots) is small (top and bottom targets on left). The ideal situation is for both bias and imprecision to be low (bottom left target).

As we said earlier, random errors are usually assumed to exhibit a normal or Gausssian distribution. This property allow us to make probability statements about measurement accuracy. Figure 9-3 illustrates this concept.

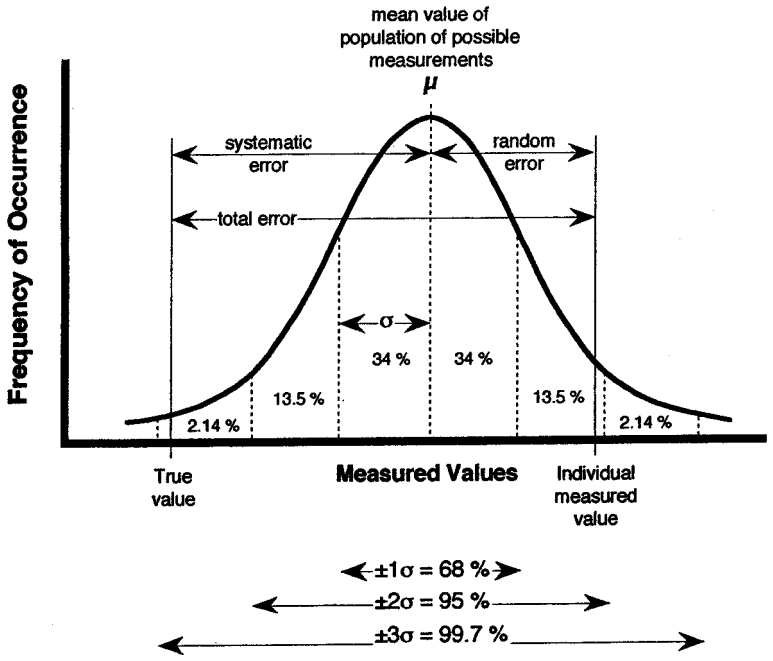


Figure 9-3. Measured values expressed in the form of a Gaussian frequency distribution. The difference between the true value and an individual measured value is the sum of both systematic and random errors. The random errors are what make the measured values appear Gaussian. That is, most random errors are small and clustered near the mean value, μ. The assumption of a Gaussian distribution allows us to predict, for example, that 95% of measurements will lie within plus or minus 2 standard deviations, σ, of the mean. See the chapter on Basic Statistical Concepts for a more detailed explanation of frequency distributions.

Linearity

A device is linear if a plot of the measured values versus the true values can be fitted with a straight line. For a linear device, the ratio of the output to the input (referred to as the *static sensitivity*) remains constant over the operating range of the measurement device. Linearity is desirable, because once the system is calibrated with at least one known input, unknown input values will be accurately measured over the linear range.

The linearity (or rather nonlinearity) specification for a system can be assessed by first fitting the best straight line to the devices output (measured) values over the range of acceptable input values. The "best" straight line is determined by using the statistical procedure known as least squares regression analysis. The resulting line will have the form of $Y = a + bX$. The value Y in the equation is the estimated mean value of repeated measurements of the true value of X. The parameter a in the equation is the estimate for the y intercept (the point where the regression line crosses the Y axis in the plot). The parameter b is the slope of the line (slope = the change in Y for a given change in X). Together, they give estimates of constant and proportional systematic errors respectively.

The linearity specification for an instrument is usually defined as the maximum observed deviation (vertical distance from the measured value to the line) from the regression line expressed as a percentage of full scale or of the reading, similar to accuracy (Figure 9-4). Remember that the linearity specification is relative to the best straight line through the data, while accuracy is relative to the line of identity. For a device with negligible systematic error, the specification for linearity is equivalent to a specification for accuracy, because the straight line that best fits the data is the line of identity. Thus, some commercial instruments give only a linearity specification and not an accuracy specification. On the other hand, an accuracy specification but not a linearity specification may be given if linear behavior of the device is implied by a fixed sensitivity specification.

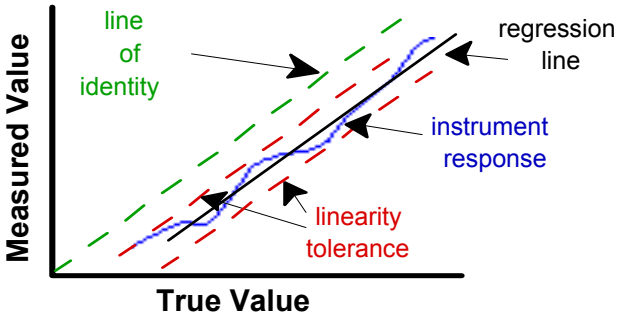


Figure 9-4. Illustration of linearity specification.

Calibration

Calibration is the process of adjusting the output of a device to match a known input so that the systematic error is minimized. *Calibration verification* is the process of measuring a known value with a calibrated device and making a judgment of whether or not the observed error is acceptable for future measurements.

For a linear measurement system, calibration can be a simple two-step procedure. First the readout is adjusted to read zero while no input signal is applied to the instrument. (A modification of this procedure is to select a known input having a low value on the scale, such as the use of 21% oxygen during the calibration of an oxygen analyzer.) Next, the sensitivity (also called gain or slope) is set by applying a known value at the upper end of the scale (such as 100% oxygen for an oxygen analyzer) and adjusting the readout to this value (Figure 9-5). If the instrument has good linearity, the readouts for all inputs between these two points will be accurate. If the instrument is not very linear, and it will be used for only a part of its measurement range, it should be calibrated for just that range. That is, using two points that at the upper and lower range of values that the instrument will be used to measure.

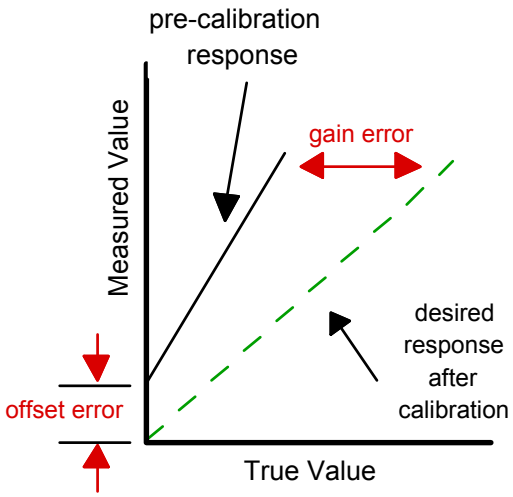


Figure 9-5. The two-point calibration procedure. First the offset is adjusted, then the gain (sensitivity) is corrected. For example, when applied to a flow sensor, the offset error would be corrected by occluding the meter so there was no flow and adjusting the readout to zero. Then, the sensor is exposed to a known flow, for example 10 L/min, and the readout is adjusted (using a separate gain control) to read 10 L/min.

Sources of Bias (Systematic Error)

Constant Error: If the zero point is not set correctly but the gain is correct, the instrument will be biased. The readings will be low or high over the entire scale (Figure 9-6A). This is also referred to as *offset error*. *Drift error* is a form of time dependent offset error in which the changes occur over time.

Proportional Error: If the zero point is set correctly but the gain is wrong, bias will be dependent on (ie, proportional to) the input level. The higher the true input value, the more error there is in the measured value (Figure 9-6B).

Range Error: Range error occurs when the true value of the input signal is outside the operating range of the instrument (Figure 9-6C). Signals that are either below or above the calibrated scale values will be clipped (the true value changes but the readout does not). In the worst case, the instrument may be damaged when exposed to over-range conditions.

Hysteresis: If an instrument gives a different reading for a given input value depending upon whether the input is increasing or decreasing, the device is said to show hysteresis (Figure 9-6D).

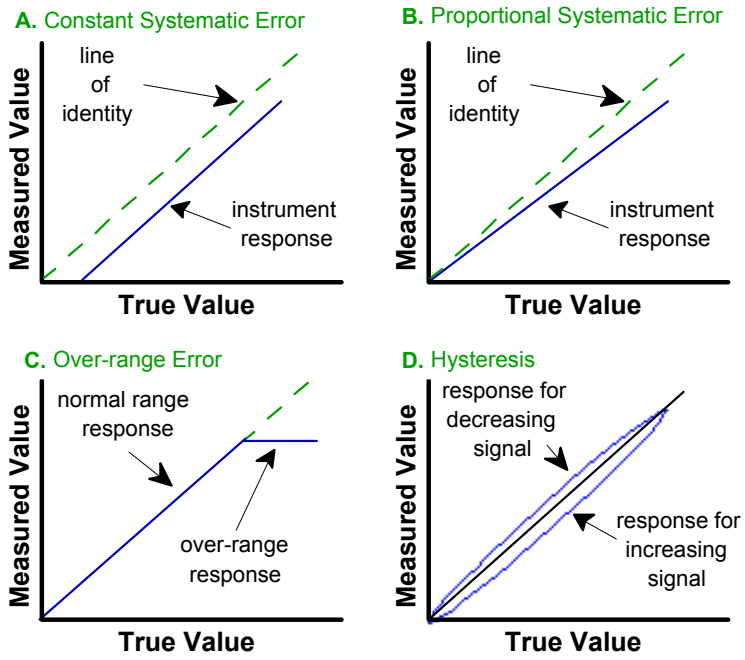
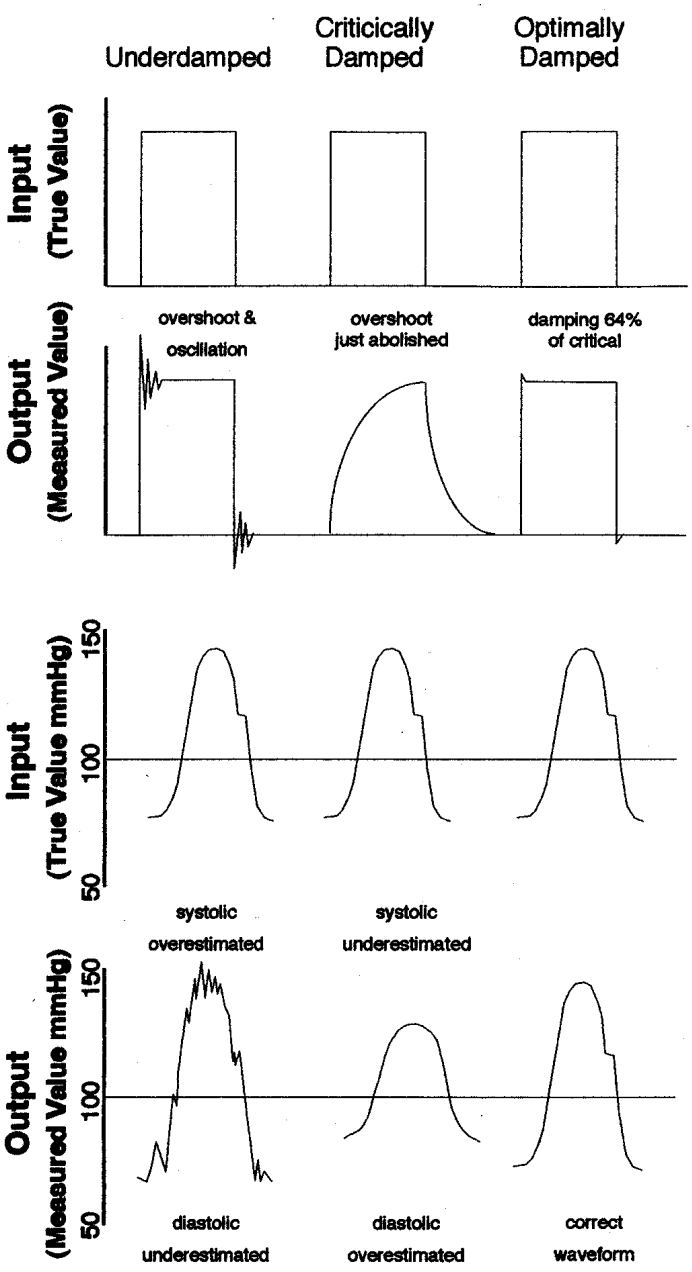


Figure 9-6. Common sources of measurement error. **A**. Constant systematic error in which the instrument reading is always higher than the true value. **B**. Proportional systematic error in which the instrument reading is higher than the true value and the error gets larger as the reading gets higher. **C**. Range error can occur when measurements are made outside the instrument's calibration level, resulting in "clipping" the signal. The example shown is an over-range error. Beyond the instrument's highest level the readout may stay constant even though the true value increases. **D**. Hysteresis, where the instrument reads too low as the signal increases and too high as the signal decreases.

Response Time. Response time is a measure of how long it takes a device to respond to a step change (ie, and instantaneous change from one constant value to another) in the input. There are two accepted methods for stating response time. The first is to simply give the *time constant*, which is the time required for a device to read 63% of the step change. For example, if an oxygen analyzer giving a stable reading in room air is suddenly exposed to 100% oxygen, the time constant is the time required for the meter to read 50% (0.63 x [100-21] ≈ 50). Alternatively, response can be expressed as the time necessary to reach 90% of the step change (sometimes modified as the time to go from 10 to 90%). For example, a 90% response time of about 100 ms is required for breath-by-breath analysis of respiratory gas concentrations. Slow response times can cause errors during calibration if you don't allow enough time for the instrument to stabilize at the known values. For practical purposes, it takes about 5 time constants to reach a steady state value. The relation among measured value, time, and the time constant is given by:

$$measured\ value = 100(1 - e^{-t/\tau})$$

where the measured value is expressed as a percentage of the steady state value, e is the base of natural logarithms (approximately equal to 2.72), t is time, and τ is the time constant expressed in units of time (the same units as t).



Frequency Response: Frequency response is a measure of an instrument's ability to accurately measure an oscillating signal. Measurements will generally either underestimate (attenuate) or overestimate (amplify) the true signal amplitude as the frequency increases (Figure 9-7). A system will generally follow an oscillating signal faithfully at low frequencies but amplify the signal as the frequency increases due to *resonance*. Resonance is a property of any oscillating mechanical system that has both inertia and elastance. It occurs when the potential and kinetic energies of these components are stored and released in synchrony (as in a pendulum). The frequency at which this occurs is called the resonant frequency or natural frequency. At higher frequencies, the system will attenuate the signal.

Figure 9-7. The top two graphs show the response to a square wave input of three different blood pressure transducers with different damping. The bottom two graphs show how damping affects the measurement of blood pressure.

It can be shown mathematically (through Fourier analysis) that any complex signal waveform can be constructed by combining sine waves at different frequencies, amplitudes, and phases (Fig. 9-8).

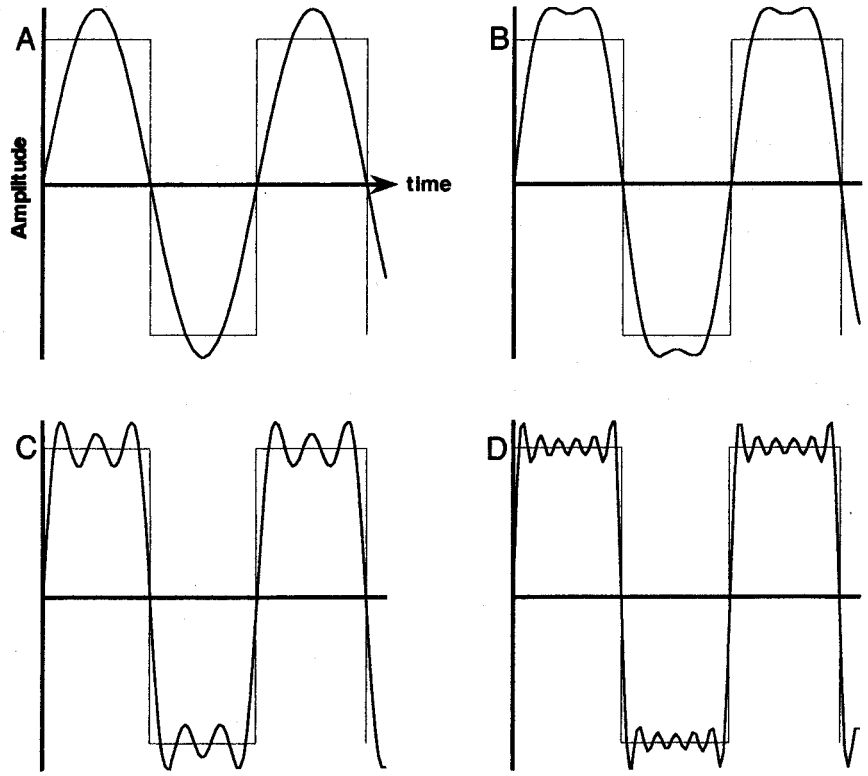


Figure 9-8. A complex signal waveform can be constructed by combining simple sine waves. In this example, a rectangular waveform is "built up" by adding harmonics of different amplitudes. Harmonics are sinusoids whose frequencies are multiples of the main or fundamental frequency. The amplitudes of the harmonics decrease as their frequency increases. **A**. The rectangular waveform is first approximated by a sine wave at the same frequency (ie, the first harmonic). **B**. Summation of the first and third harmonics. **C**. Summation of first, third, and fifth harmonics. **D**. Summation of the first, third, fifth, ninth, and eleventh harmonics. As more harmonics are added, the waveform becomes more rectangular. On the other hand, if a rectangular waveform is damped, it becomes more rounded as harmonics are filtered out.

A system is said to be *damped* when some of the signal component frequencies are attenuated. A *critically damped* system is one that follows a step-change input with maximum velocity but does not overshoot (Fig. 9-7). An *optimally damped* system will measure all signal frequencies within the working range with equal amplitude. Such a system is said to have a "flat" response, meaning that the amplitude distortion is less than $\pm$ 2 percent up to 66 percent of the undamped resonant frequency. A device that is tuned for a given frequency range may exhibit errors in the magnitude and timing of the measured signal if used at higher frequencies. Frequency response problems are especially evident for pressure and flow measurements and with instruments like analog meter readouts and strip chart recorders.

Loading Error. A basic axiom of measurement theory is that the measurement process inevitably alters the characteristics of the measured quantity. Therefore, some measurement error will always be present. For example, placing a pneumotachometer in a flow stream changes the flow rate because of the added

resistance. Also, when electronic devices are coupled, unrecognized electronic loading can occur and can be quite serious.

Environmental Conditions. If a measurement system is used under significantly different conditions (like pressure or temperature) than those under which it is calibrated and if no correction is made, systematic errors result. A typical example is the effect of barometric pressure, humidity, or anesthetic gases on polarographic oxygen analyzers.

Operator Errors. Between-observer variations in measurement technique and within-observer habits (such as always holding your head to one side while reading a needle and scale having parallax) can result in bias. Human observers also exhibit what is known as "digit preference". Anytime an observer must read a scale, a guess must be made as to the last digit of the measurement. Most people tend to prefer some digits over others. For example, a blood pressure of 117/89 torr is rarely recorded. Observers tend to prefer terminal digits of 0 and 5. Thus, readings such as 120/85 torr are far more commonly recorded. The way observers round numbers and select significant digits can also introduce error.

Sources of Imprecision (Random Error)

Noise. All measurements are subject to some degree of minor, rapidly changing disturbance caused by a variety of environmental factors. This is called noise. It may be difficult to trace and is not reduced by calibration. Noise distortions, however, are usually considered to occur randomly. That means their effects should cancel out if enough repeated measurements are made. Noise can be particularly disturbing with weak signals that are highly amplified. The noise is amplified along with the signal, such that a limit is eventually placed on the sensitivity of the measurement. For example, an electrocardiograpic (ECG) signal may be contaminated with intercostals electromyographic (EMG) signals of equal amplitude along with noise form electrochemical activity between the skin and the electrode. On the way to the amplifier, the signal is subject to electrostatic and electromagnetic noise (at 60 Hz) from nearby power lines. Radio frequency noise may be added from surgical diathermy or radio transmitters. Physical movement of the electrode cable changes its capacitance and may add low-frequency noise. The most common form of electrical noise is called *thermal* or *Johnson noise*, caused by the random motions of electrons in conductors.

The efficiency with which a signal can be distinguished from background noise is defined as the signal-to-noise ratio, using a logarithmic bel scale. One bel is a ratio of 10:1. Two bells is 100:1. A more convenient unit is the decibel, which is on-tenth of a bel:

$$\text{decibel (dB)} = 10 \log_{10} (P_2/P_1)$$

where P_1 is input power and P_2 is output power. Thus, a ratio of 100:1 is 20 dB.

Nonlinearity. Nonlinearity is considered to cause imprecision because it will introduce an unpredictable error that varies over the operating range, depending on the level of the input signal (in contrast to proportional error that is predictable). Errors due to nonlinearity can be minimized by calibrating at two points within the range in which most measurements will be made.

Operator Errors. Human errors during measurement can also introduce imprecision. For example, variations in readings given by different observers may be caused by reading a dial at different angles, failing to judge the exact reading consistently, or slight variations in preparing transducers or samples.

MEASURING SPECIFIC VARIABLES

In this section, we will review the standard techniques for measuring some of the variables common to health care research.

Pressure

Perhaps the most fundamental measurement in respiratory mechanics is the measurement of pressure. *Absolute pressure* refers to the absolute force per unit area exerted by a fluid. *Gauge pressure* is the pressure of interest referred to local atmospheric pressure. The difference, therefore, between absolute pressure and gauge pressure is the local atmospheric pressure on that particular day. Steady-state pressures are easily measured by a variety of devices; the measurement of a pressure that varies with time is a far more complicated task.

U-Tube Manometer: This device consists of a u-shaped tube (usually glass or clear plastic) filled with a liquid (eg, water for most purposes but may be mercury for measuring higher pressures or oil for smaller pressures). There is a scale attached to measure the height of one liquid surface with respect to the other (Figure 9-9). If one leg of the manometer is attached to a source of pressure and the other leg is open to atmospheric pressure, the liquid will rise in the leg with the lower pressure and fall in the leg with the higher pressure. The distance between the levels is a measure of gauge pressure (eg, in cm H_2O or mm Hg). This device is only good for measuring static pressures and is often used to calibrate other types of pressure sensors.

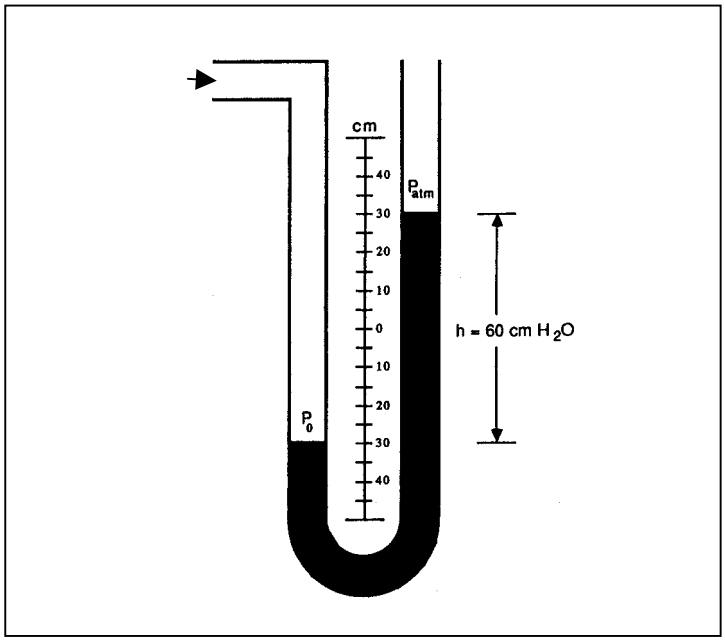


Figure 9-9. The U-tube manometer. The height, h, indicates that the measured pressure, P_0, is 60 cm H_2O above atmospheric pressure, P_{atm}.

Bourdon-Type Pressure Gauges: Bourdon-tube gauges are available for reliable for static pressure measurements. Basically, they employ a curved tube with a shape that changes under an applied pressure difference. This shape change is mechanically coupled to the rotation of a shaft; a needle attached to this shaft gives the pressure reading on a dial (Figure 9-10). The mechanical components are cumbersome and the dynamic response relatively slow compared with the usual methods for dynamic pressure measurements. This type of gauge is commonly used on compressed gas cylinders. Some mechanical ventilators use a variation of this type of gauge for measuring airway pressures. Bourdon-tube gauges tend to go out of calibration easily.

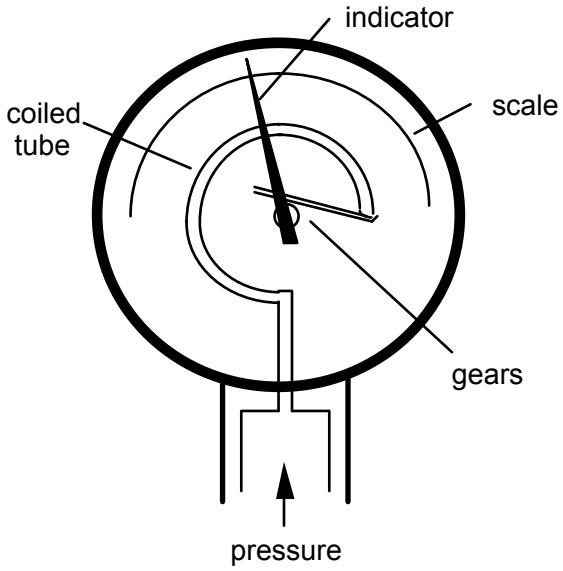


Figure 9-10. The Bourdon-tube pressure gauge.

Diaphragm Pressure Gauges. Two different arrangements exist for measuring pressures by means of detecting the deflection of a metal diaphragm under an imposed pressure loading. In one type, a thin circular metal plate with clamped edges deflects under an applied pressure loading. The engineering theories of plate deflection can be used to relate the displacement at the center of the plate to the pressure difference applied. The deflection can be measured by strain gauges bonded to the surface of the diaphragm (Figure 9-11).

The strain gauge is an element with an electrical resistance that changes as it is deformed. Continuously measuring the voltage drop across a circuit containing this resistance (the "bridge") is equivalent to continuously measuring the displacement of the diaphragm. The second method for measuring the displacement involves the fact that movement of a magnetic core between two coils changes the magnetic coupling between the coils, and therefore changes the voltage output of the secondary coil. In this case, there has to be a signal input to the primary coil-the carrier-and a means to decode the signal that appears in the secondary coil. With either of these methods, the displacements of the diaphragm are generally small and the diaphragms used are extremely thin and light. Therefore the dynamic response of the solid components of the device is very good. The frequency response, however, may be limited by the fluid in the connecting tube and in the chambers of the gauge.

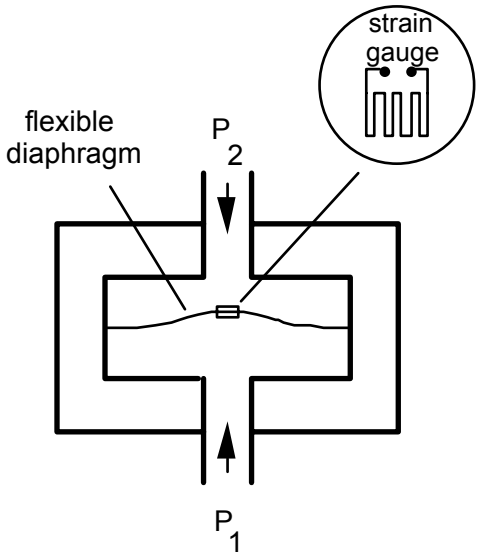


Figure 9-11. Diaphragm pressure gauges. The diaphragm is clamped at the edges and deflects under applied pressure load ($P_1 – P_2$). The magnitude of the load is detected by the elongation or compression of the strain gauge.

Piezoelectric Crystals: Very small pressure gauges can now be constructed of piezoelectric crystals. A piezoelectric substance, when stressed or deformed, produces a potential difference (voltage drop) across the crystal. Sensing this voltage therefore measures the deformation of the crystal. Because there are essentially no moving parts, this device is very good for both static and dynamic pressure measurements. Piezoelectric pressure sensors are used in most modern medical equipment, including ventilators.

Flow

Several types of flow-measuring devices are available; again, the determination of which to use depends upon whether one wants to measure steady flows or time-varying (unsteady) flows. In respiratory mechanics we are generally interested in flow into and out of the airways and in how this changes lung volume. We therefore speak of volume as if it flows-flows are expressed in liters per minute, for example. This shorthand notation overlooks an important physical fact: gases flow; volumes do not. When we speak of a *volume flow* of so many liters per minute, what we are really saying is that the mass of gas that has exited from the lung over that time would occupy a volume of so many liters *at some specific temperature and pressure.* Thus, flow measurements require accurate temperature and pressure measurements to be accurate.

In terms of the frequency response required from flow-measuring instrumentation, studies of the usual maneuvers performed in pulmonary function laboratories show that flow from humans can have significant frequency components up to 60 Hz. This approaches the limits of the available instrumentation.

Rotameters: The rotameter is commonly used to measure steady flow. In the pulmonary laboratory, it finds its greatest use in calibrating other devices. Essentially, it is a tapered vertical tube containing a float (Figure 13-12).The float rises as flow rate is increased, thereby increasing the cross-sectional area

through which flow can occur. Volume flow is then directly proportional to this open area. Because the float is stationary, its weight is then balanced by the pressure of the gas on its surfaces and the drag caused by the flow going past it. The scale is usually calibrated in terms of liters per minute of volume flow, but the specific temperature and pressure conditions that prevail on that day must be recorded.

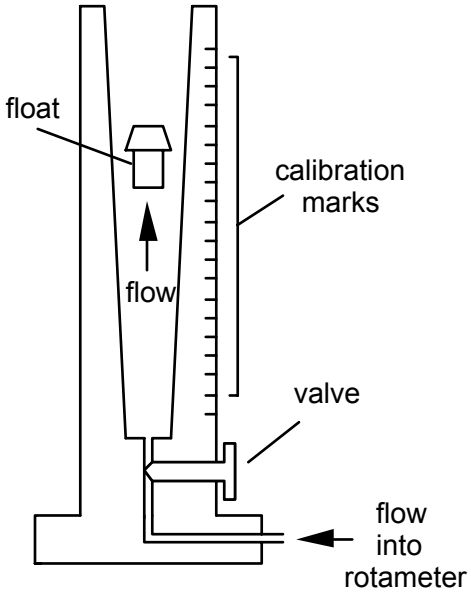


Figure 9-12. A rotameter.

Pneumotachometers. Pneumotachometers are devices designed to produce a pressure drop when exposed to a given flow. The pressure drop measured and flow is then inferred from the pressure measurement. Pneumotachometers are the workhorses of respiratory research flow measurements. They are of two basic types (Figure 9-13), the capillary bed (Fleisch) type and the screen (Silverman) type. The basic principle behind the bed of packed capillary tubes (or other close-packed channels) is that the tubes are of such small diameter that near their center the flow is essentially one-dimensional and relatively steady. The Poiseuille equation for steady flow through a pipe shows that the pressure drop is linearly related to the volume flow rate. This concept works well for steady or slowly changing flows. However, if a sinusoidal flow is generated, (for example, with a piston pump), one finds that the frequency range over which a Fleisch pneumotachometer faithfully records the flow is limited.

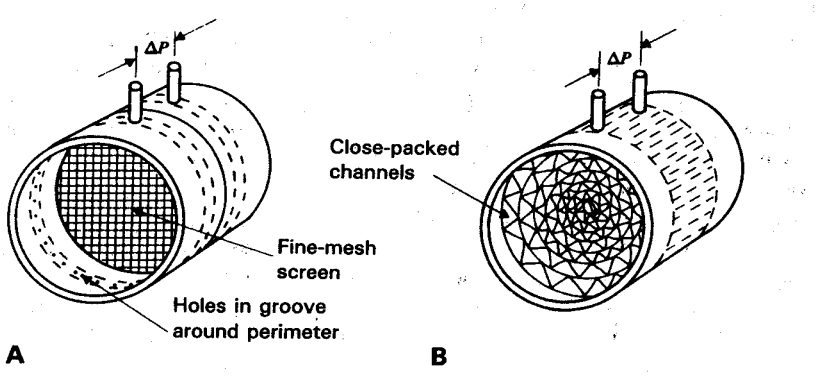


Figure 9-13. Pneumotachometers. A. Silverman or screen type. B. Fleish or capillary tube type.

Screen-type pneumotachometers rely on a fine mesh screen to produce a pressure drop related to flow through the screen. Their frequency response is better than that of the capillary type pneumotachometers, but they generate much more turbulence in the flow and hence noise in the pressure signal. Pneumotachometers may be heated to prevent condensation of exhaled water vapor and subsequent obstruction of the channels for flow.

A modification of the Silverman pneumotach involves the use of a solid membrane of plastic instead of a screen. There is a cut in the membrane creating a flap. The flap deforms and creates a whole in the membrane as flow passes through the pneumotach. This design is often used to make inexpensive, disposable sensors that are not affected by condensation (such as in ventilator circuits), which would ruin the calibration of a screen pneumotach.

Hot Wire Methods: This type of flow detector depends on the principles of hot wire anemometry. A heated wire is placed across the channel of a cylinder, much like that of a pneumotachometer. Electrical heating is supplied and either the wire temperature or temperature of the flowing gas is monitored. Flow across the wire produces a cooling effect. The degree of cooling can be measured and calibrated to provide an indicator of flow. Because the heat is taken up by molecules of the gas, this is a direct measurement of *mass flow* rather than volume flow.

Vane Transducers: For relatively steady flows, the displacement of a mechanical device such as a rotor or a disk may be used to measure flow (Fig. 9-14). In the device shown, the gas strikes the vanes and a fixed quantity of gas is contained in each section, much like people passing through a revolving door. As the vane turns, the gas is delivered to the outlet; the rotations are counted and the appropriate flow rate is displayed on some scale. In this case, the rate measured is the volume occupied by the gas; the rate of mass flow depends upon the density (and therefore the composition; temperature, and pressure) of the gas. These devices are not used for the measurement of flows that vary rapidly in time (such as during a forced vital capacity maneuver). However, they are commonly used for such things as measuring spontaneous tidal volumes on patients at the bedside.

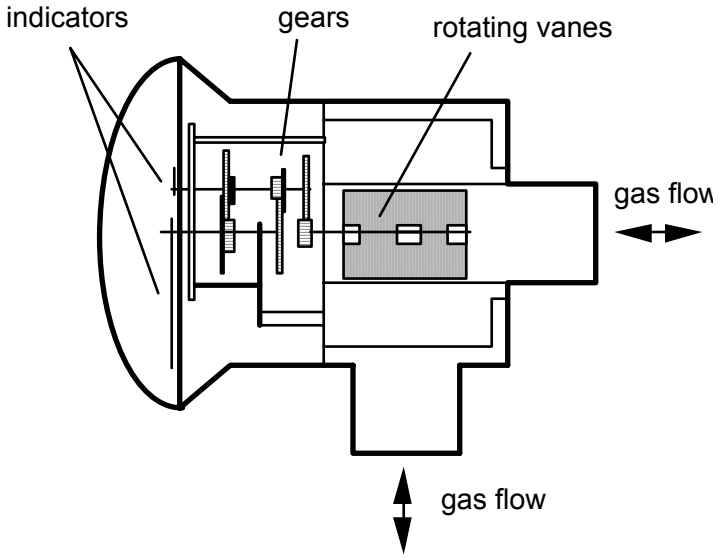


Figure 9-14. A typical vane flow transducer. The example shown is a Wright Respirometer.

Turbine Flowmeters. For these meters the float is replaced by a vaned disk or turbine wheel. The viscosity or drag generated by the flow spins the wheel. A magnetic probe in the body senses the number of vanes rotating by in a given time much like the cruise control detects engine rpm and therefore vehicle speed in an automobile. These meters are somewhat better in their transient response than those discussed earlier and occasionally are used to measure time-varying flows. Again, the relationship between rotational speed and mass flow depends on gas viscosity and density, and hence on local temperature and pressure.

Ultrasonic Transducers. Ultrasonic flowmeters consist of two piezoelectric transducers separated by a known distance, and electrical circuitry to generate and then detect high-frequency sound waves. In some designs, the transducers are mounted axially. In others, the wave crosses the flow at a shallow angle to the axis. Each transducer is alternately used as a transmitter and then a receiver, and the time of transit for the sound wave measured. Waves traveling in the direction of fluid flow arrive more quickly than sound waves traveling against the flow. From the differences in transit time the flow velocity can be computed. In general, these devices are capable of use in unsteady flow conditions. Their signal at zero flow (the baseline), however, tends to drift over time, making them relatively unreliable.

Volume

Spirometer: One of the first, and the simplest, methods to track changes in lung volume is to collect gases in a spirometer, an inverted, counterbalanced can with a water seal (Figure 9-15). Because the spirometer is at a different temperature than the lungs, the spirometer volume changes must be corrected for temperature to represent changes in lung volume. The spirometer gives only a relative measure of volume in that it detects *volume changes.* In addition, because of the mechanics of the spirometer and the mechanical linkages to the drum and pen on which volume change is recorded, its dynamic response is not optimal. In fact, the frequency response deteriorates at around 4 Hz.

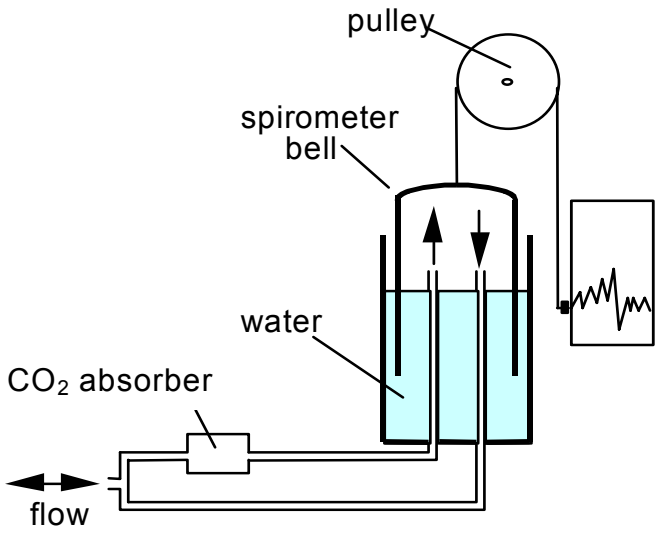


Figure 9-15. Water seal spirometer

Tracer Gas: Methods for determining *absolute volume* frequently make use of tracer gases. An inert gas (such as helium) not usually present in the respiratory system and not highly soluble in blood can be

introduced to a closed system such as a spirometer. The amount introduced is known. If the subject respires the gas mixture until a uniform concentration exists in the closed system, conservation of mass together with a measurement of the concentration gives the volume of the subject's lungs *that exchanges gas with the system:* any blebs or obstructed areas that contain air but do not receive the gas are not included in this measurement.

A variant of this process employs pure oxygen in the closed system and uses the nitrogen present in the subject's lungs as the tracer gas. This is known as the nitrogen washout method, and again measures only the volume of gas in the subject that exchanges gases with the environment.

Plethysmograph: Most of the literature of traditional respiratory mechanics fails to distinguish between the volume flow of gas out of the lung and the rate of change of lung volume. Consider a simple syringe whose outlet is closed by a stopcock. Pressure on the plunger will change the volume of gas in the syringe because of gas compression; however, no gas flows out of the syringe. If the stopcock is opened a crack, and if one pushes slowly and gently on the plunger, gas will flow without much change in the pressure in the syringe, that is, without much gas compression. The volume of gas exiting the syringe will be very close to the volume change in the syringe itself . If instead one attempts to force the plunger, gas compression occurs, and the volume in the syringe decreases to a greater amount than can be explained by the amount of air that has exited. In fact, this gas compression makes possible a measurement of lung volume by a process called body plethysmography. You may gain some insight into how plethysmographs function by a simplified discussion employing the relationships among pressure, temperature, and density of ideal gases.

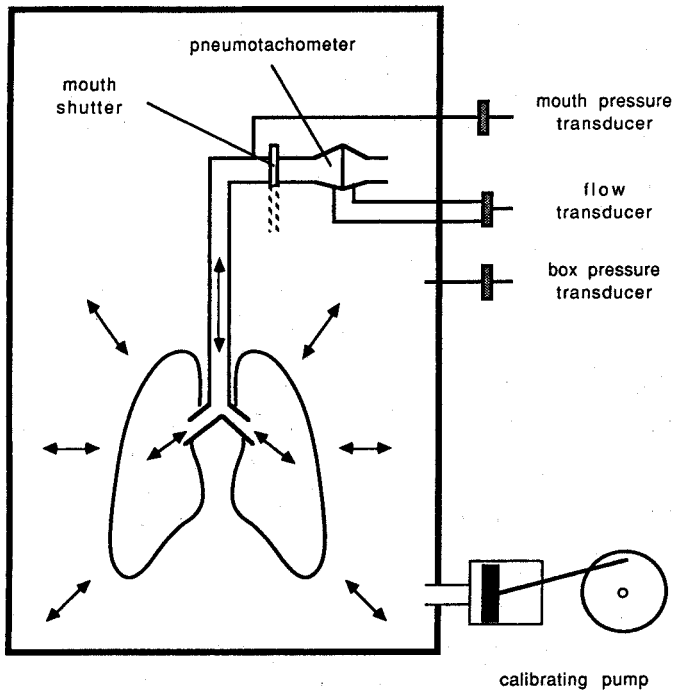


Figure 9-16. Pressure type plethysmograph.

In the most common type of plethysmograph, the pressure plethysmograph (Figure 9-16), there is a rigid box around the subject, a mouthpiece to breathe through, and pressure taps at the mouth end of the mouthpiece and in the box itself. The box is attached to a small piston pump with a known stroke

volume. In addition, there is a shutter at the subject's mouth that can be electronically closed. In the diagram, we assume that the lung behaves as a single compartment, that is, that the pressure is the same everywhere within the lung. When the gases in the subject and in the box are at the same constant temperature, the gases will obey Boyle's Law:

$$PV = \text{constant}$$

Furthermore, the total volume inside the box is constant, so that small changes in lung volume (dV_L) and box volume (dV_B) must be equal and opposite:

$$dV_L = -dV_B$$

When the shutter is closed and the gas in the subject's lungs is compressed and expanded by panting against the closed shutter, changes in pressure and volume occur in both the lung (P_L, V_L). and in the box (P_B, V_B). These changes must be related by:

$$\frac{dP_B}{dP_L} = \frac{P_B}{V_B} \cdot \frac{V_L}{P_L}$$

If one has previously quantified the relationship between changes in box pressure and volume by changing the box volume with the piston pump, the calibration factor (P_B/V_B) is known. Thus one can measure the absolute volume of the lung by measuring pressure and its changes in the lung (that is, mouth pressure) and box during the occlusion maneuver.

Volume (V) measurements are commonly derived from flow $(\dot{V})$ measurement. This is possible because, mathematically, volume can be expressed as the integral of flow between two point in time (t):

$$V = \int \dot{V}\, dt$$

In the simplest case, if flow is constant then volume = flow x time. Conversely, flow is the derivative of volume:

$$\dot{V} = \frac{dV}{dt}$$

Again, the simplest case is when flow is constant such that volume is simply the change in volume divided by the change in time.

Volume measurements can also be derived from only pressure measurements in some cases. Ventilators are often evaluated by connecting them to a simulated lung consisting of simply a rigid-walled container. As gas enters the container from the ventilator, the pressure inside the container rises. The internal energy also increases so that the temperature of the gas increases. If the container is a perfect insulator, the gas will not lose heat during the process and an *adiabatic* compression is said to occur. The pressure change is a function of the quantity of gas added (measured in moles) and the change in temperature. Therefore, a simulated volume measurement can be made by measuring the pressure inside the container. The equation relating volume and pressure for an adiabatic compression is:

$$V = P\left(\frac{V_c}{1.9P_B}\right)$$

where

V = simulated volume measurement, which proportional to the pressure measurement (L)

P = pressure change, above ambient barometric pressure, inside the container (cm H_2O)

V_c = fixed volume of container (L)

P_B = barometric pressure (mm Hg)

Sometimes the container is filled with fine copper wool (very thin copper wire, densely packed). This material provides a huge surface area for heat exchange and acts like a thermal buffer. The effect is that the temperature inside the container remains relatively constant as pressure increases. This process is known as an *isothermal* compression. The equation relating volume and pressure for an isothermal compression is:

$$V = P\left(\frac{V_c}{1.36P_B}\right)$$

In practice, a pressure transducer is connected to the container and the output calibrated to read volume using a calibrated syringe to inject known volumes into the container. The fixed volume of the container can be selected to simulate adult, pediatric, and neonatal patients (Table 9-1). Note both of the equations above can be solved for *V/P*, which is the simulated compliance of the lung model.

Table 9-1. Lung model parameters recommended by the American Society for Testing and Materials (ASTM F1100-90, published 1990).

Simulated Patient	Compliance (mL/cm H_2O)	Volume* (L)	Resistance (cm H_2O/L/s)	Flow** (L/s)
Adult	50	51.6	5	0 – 2.0
Adult/pediatric	20	20.6	20	0 – 1.0
Pediatric	10	10.3	50	0 – 0.5
Pediatric/neonatal	3	3.09	200	0 – 0.1
Neonatal	1	1.03	500	0 – 0.75

*Volume of a rigid-walled container required to achieve specified compliance when filled with wire wool (2% by volume) at normal atmospheric pressure (760 mm Hg). Changes in atmospheric pressure will change the simulated compliance of the container so that some method of adjusting volume may be necessary to keep within the specified tolerance for compliance of ± 5%.

**Range of flows for which linear resistance is calculated.

Humidity

In some experiments we need to measure the amount of water vapor present in a gas sample. Of all the physical quantities measured in a pulmonary function laboratory, humidity is probably the most difficult to measure accurately. Errors on the order of 2% are at the leading edge of the available technology.

The "gold standard" is the absorption method, by which the gas sample of interest is introduced into a chamber containing a desiccant and the amount of water extracted is determined by the change in weight. This method is not, however, rapid or practical, particularly for measurements over time. To perform repeated experiments at reasonable time intervals, some indirect measure of humidity must be used. For serial measurements at fairly large time intervals, the dew point hygrometer has long sufficed. In this method, the gas is cooled while its pressure is maintained constant. Some method must be provided for determining the temperature at which condensation first appears. This may involve a polished metal mirror that is visually observed, or in a sophisticated hygrometer may be a quartz crystal oscillated at its resonant frequency. In the latter case, as water condenses on the crystal, its weight increases and the resonant frequency changes. This change can be detected. The temperature at which the condensate forms is related to, but generally not equal to, the dew point temperature of the air. The dew point in turn is a measure of humidity in that it is the temperature at which the amount of water vapor present fully saturates the volume of air in contact with it. The discrepancy between the condensation and dew point temperatures occurs because the appearance of the droplets is affected by the nature of the surface, any contaminants that may form nuclei for condensation, and other factors.

Perhaps the most practical methods for determining humidity are those based on capacitance changes. A capacitor is a circuit device consisting of two metal plates separated by a dielectric layer that acts as an insulator. When the two plates are electrically charged, a voltage difference is produced that depends upon how well the dielectric material avoids becoming polarized. In capacitance humidity sensors, the dielectric material is one that can absorb water. Because water polarizes so readily this changes the dielectric constant of the sandwiched material and therefore changes the output of the sensor. Presently, the best sensors have employed a polymer as the dielectric material. Unfortunately, at high relative humidity such as that encountered in the respiratory system, the sensors tend to be unstable. In addition, the sensors are very sensitive to temperature changes and the results must be correlated with local temperature. Thus, truly dynamic humidity measurements are still difficult to make.

Humidity sensors are calibrated by exposing them to the air above a saturated salt solution. This is because the relative humidity above the solutions can be predicted (Table 9-2).

Table 9-2. Relative humidity (%) in air above saturated salt solutions as a function of solution temperature.

	20° C	**25° C**	**30° C**	**35° C**
Lithium chloride (LiCl)	12.4	12.0	11.8	11.7
Sodium chloride (NaCl)	75.5	75.8	75.6	75.5
Potassium sulfate (K_2SO_4)	97.2	96.9	96.6	96.4

One procedure for calibration is a follows:

1. Select a salt that will provide the desired humidity calibration level). LiCl salt will provide a low calibration point, NaCl (table salt) will provide a midrange calibration point, and K_2SO_4 will provide the high calibration point. Pour approximately 20 mL dry salt into a 100 mL beaker. Add enough pure distilled water (do not use deionized water) to just cover the salt (no standing water above the salt).
2. Stir the moistened salt.
3. Cover the beaker with plastic and allow to sit for at least 24 hours.
4. Check that there is some undissolved salt with a clear saturated salt solution above it. If not, add more water and stir. If all the salt is dissolved (because too much water was added initially) add more salt and wait another 24 hours.
5. Place the humidity sensor in the beaker approximately 20 mm above the solution. Do not let the sensor touch the solution.
6. Cover the beaker, making sure to seal tightly around the cable. Let stand for about 2 hours so that the humidity level reaches a steady state.
7. Note the room temperature and calibrate the output of the humidity sensor according to Table 9-2.
8. Repeat steps 5 through 7 to get a calibration at three points using the three salts.

SIGNAL PROCESSING

The transducers mentioned all provide some sort of signal together with some noise. Some of them, such as the inductance-coupled pressure transducer or the strain gauge transducer, require some electrical input to be operated upon to produce an electrical output. All of these signals must be processed in some fashion to be useful. For many of the mechanical meters, processing may be done internally in terms of the connections inside the meter to pointers and scales. This internal mechanical processing may limit the frequency response or damp out unwanted or excessive oscillations. Thus, not all signal processing is electrical.

Several particularly prevalent methods of signal processing need to be mentioned. The first is amplification. Small signals often need to be boosted to be used by other electrical equipment. On the other hand, some very large signals occasionally need to be reduced. Both of these processes, oddly enough, are referred to as *amplification.* The *gain* is the ratio of the output signal to the input signal amplitude. If the gain is greater than 1, we have what we traditionally think of as amplification. If the gain is less than 1, the device is still said to be an amplifier, but it really is an attenuator. Given what was said about Fourier analysis, we obviously want the gain to be constant over the frequency range of interest. Just as importantly, amplification should not introduce a phase shift into the signal.

Signals also can be filtered. A *filter* is a device that removes certain frequency components from the signal. If it removes high-frequency components, it is called a low-pass filter. If it removes low-frequency components, it is a high-pass filter. If it removes low and high components, it is referred to as a band-pass filter. Filtering can be accomplished either electronically or mathematically from data that have been taken without a filter. Generally, pulmonary function laboratories use low-pass filters to get rid of noise and other signals at high frequency. This makes sense because if most of our devices distort signals at higher frequency, it is just as well to remove these frequencies from the signals we are trying

to analyze. On the other hand, signals of interest totally unknown to us may be present in the frequencies we are eliminating. The frequency at which the filter starts to attenuate the signal by a given amount is called the *corner frequency,* and is generally specified on the filter. Another characteristic of importance is the *roll-off,* or the rate of attenuation. It is generally specified in terms of decibels (db) per decade. The higher the roll-off, the more effective the filter is in removing frequencies near the corner frequency. This assures that these frequencies will be nearly totally eliminated rather than merely phase shifted and somewhat reduced in amplitude. The names of the types of filters (Butterworth, Bessel, Chebyshev) refer to the mathematical relationship between amount of attenuation and frequency.

RECORDING AND DISPLAY DEVICES

An oscilloscope is a cathode ray tube on which a voltage can be displayed versus time or versus a second voltage. For monitoring purposes, signals may be observed on the oscilloscope and may even be photographed on its face to provide a permanent record. Alternatively, the signal may be sent directly to a strip chart recorder or to a plotter. However, the dynamics of the pen systems will generally preclude accurate recording of signals that vary rapidly in time.

Signals may also be recorded by computer. In this case, the sampling rate at which the computer acquires the data is important. A continuous signal presented to a computer is recorded at fixed increments in time the sampling interval. This is called *analog to digital conversion* because a signal continuous in time (analog signal) is represented in the computer by the discrete sample values obtained at certain time points (digital representation). This conversion is not as straightforward as it seems. In particular, if the sampling rate is too slow a phenomenon known as *aliasing* occurs. The higher-frequency components of the signal are shifted to appear as lower-frequency components and the representation of the signal is consequently distorted. In order to avoid aliasing, the sampling interval must be less than half the period for the highest frequency component of the signal. Thus, if the signal is expected to have a component of 50 Hz, one must sample at least 100 samples per second. The software may or may not support a sampling rate fast enough for the experimenter's purposes. In addition, the converters in various computers differ in the range of input voltage expected and in how these signals are represented in the computer's memory. These details are important for any given application.

Perhaps the best way to set up a laboratory for making measurements is to purchase a computer-based system that combines sensors with signal processing hardware and special software makeing it easy to capture, view, and analyze measurement data. One of the most popular systems is called LabVIEW by National Instruments Inc. You can learn more about this product at www.labview.com

QUESTIONS

Definitions

- Systematic errors
- Random errors
- Accuracy
- Precision
- Resolution

- Calibration
- Calibration verification

True or False

1. A U-tube manometer is used for measuring humidity.
2. A pneumotachometer is used to measure gas flow.
3. A rotameter is used to calibrate pressure sensors.
4. A hot-wire anemometer senses changes in fluid volume.
5. A spirometer is used to measure changes in gas volume.
6. A dew-point hygrometer is used to determine the humidity of a gas sample.
7. Calibration reduces systematic errors but not random errors.

Multiple Choice

1. The ideal situation for a measurement is to have:
    a. High bias and low imprecision
    b. Low imprecision and low bias
    c. High imprecision and high bias
    d. Low bias and high imprecision
2. All of the following are true about a highly linear measurement device except:
    a. The ratio of the output to the input remains constant over the operating range.
    b. The linearity specification can be assessed using least squares regression.
    c. The linearity specification is defined as the maximum deviation from the regression line.
    d. Linearity is desirable because the system does not have to be calibrated.
3. The two-point calibration procedure involves:
    a. Adjusting the sensitivity and the linearity.
    b. Adjusting the offset and the gain.
    c. Measuring a known quantity and reducing the random error.
    d. Decreasing both bias and imprecision.
4. Which of the following sources of bias occur if the zero point is not set correctly:
    a. Constant error
    b. Proportional error
    c. Range error
    d. Hysteresis
    e. Frequency response

f. Noise

5. Which source of error can be caused by electromagnetic radiation?
    a. Constant error
    b. Proportional error
    c. Range error
    d. Hysteresis
    e. Frequency response
    f. Noise
6. Which source of error is due to an improperly set sensitivity?
    a. Constant error
    b. Proportional error
    c. Range error
    d. Hysteresis
    e. Frequency response
    f. Noise
7. What type of error would you expect if the needle on the measurement device went off the scale?
    a. Constant error
    b. Proportional error
    c. Range error
    d. Hysteresis
    e. Frequency response
    f. Noise
8. If the instrument gives a different reading for a given input value depending upon whether the input is increasing or decreasing, the device is said to show:
    a. Constant error
    b. Proportional error
    c. Range error
    d. Hysteresis
    e. Frequency response
    f. Noise
9. If the signal is moving faster than the measuring device is capable of following, what type of error would you expect?
    a. Constant error

b. Proportional error
c. Range error
d. Hysteresis
e. Frequency response
f. Noise

SECTION IV ANALYZING THE DATA

Chapter 10. Basic Statistical Concepts

Statistical analysis of the numbers generated by observation and measurement in a research study is usually one of the most intimidating topics for an investigator. Unfortunately, inappropriate statistical analyses sometimes appear in published studies. The goal of this chapter is to provide a clear overview of the meaning and use of common statistical techniques. Emphasis is on the concepts and correct use more than the calculation of statistics. For this reason, simplified data sets are used, and calculations are included only to the extent necessary to discuss the concepts. The purpose of this chapter is to help you to be an educated consumer of research publications first, and to help you become a researcher second. However, this chapter and the next will give you enough information to understand published research articles and even to perform some basic statistical analyses on a computer. Beyond that, you really need to study formal statistical textbooks if you want to understand the theory.

PRELIMINARY CONCEPTS

Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data. *Biostatistics* is the application of statistics to the fields of medicine and biology. Historically, the term *statistic* referred to numbers derived from affairs of state, such as number of births or deaths, or average age expectancies. Such numbers form the basis for what are called *descriptive statistics* of today, which include frequency distributions, measures of central tendency (mean, median, and mode), measures of variability (variance, standard deviation, and coefficient of variation), standard scores, and various correlation coefficients. In the mid-1600s the *theory of probability* was developed and applied to games of chance by Pascal and Bernoulli. In 1812, the theory of probability was applied to the distribution of errors of observation by LaPlace, and later by Quetelet, who used probability models to describe social and biological phenomena, for example height or chest circumference distributions. The application of probability theory to descriptive statistics resulted in *inferential statistics.* This area of statistics allows us to test hypotheses; that is, we can obtain the probability of getting the observed results by random chance and decide whether the treatment made a difference. Inferential statistics include parametric statistics (those based on specific distributions such as the normal and t-distributions) and non-parametric tests (those that do not assume the data are distributed in any particular fashion).

Definition of Terms

Population: a collection of data or objects (usually infinite or otherwise not measurable) that describes some phenomenon of interest.

Sample: a subset of a population that is accessible for measurement.

Variable: a characteristic or entity that can take on different values. Examples are temperature, class size, and political party.

Qualitative variable: a categorical variable not placed on a meaningful number scale. An example is the variable gender, which generally has two values, male and female. Any assignment of numbers to these two values is arbitrary and arithmetically meaningless.

Quantitative variable: one that is measurable using a meaningful scale of numbers. Both temperature and class size are quantitative variables.

Discrete variable: a quantitative variable with gaps or interruptions in the values it may assume. An example is any *integer* variable, such as class size, number of hospital beds, or number of children per family. We have 1 child, not 1.5 or 1.82.

Continuous variable: a quantitative variable that can take on any value, including fractional ones, in a given range of values. An example is temperature or pH. Between a pH of 7.41 and 7.42, we can have 7.411, or 7.4115. Every value is theoretically possible, and is limited by instrumentation or application.

The distinctions between discrete and continuous variables, as well as between quantitative and qualitative, are fundamental to a presentation of statistics. Such distinctions determine the appropriateness of particular statistics, in both descriptive and inferential statistics. Their meaning will be further clarified by their use in applications.

Levels of Measurement

Statistics involves the manipulation and graphic presentation of numbers. The process of assigning numbers to things (variables) is termed *measurement,* and the numbers that result are termed *data* (singular: datum). However, the arithmetic properties of numbers may not apply to the variables measured. The problem can be illustrated with some examples. Numbers (0, 1, 2, ...) have the following properties:

Distinguishability: 0, 1, 2, and so on, are different numbers.

Ranking (greater than or less than): 1 is less than 2. If 1 is less than 2, and 2 is less than 3, is less than 3.

Equal intervals: Between 1 and 2, we assume the same distance as between 3 and 4. Therefore, 2 -1 = 4 -3.

When numbers are applied to variables, these properties may or may not hold. For example, consider the qualitative variable gender, which can be assigned a 1 for male and a 2 for female. The only numerical property that applies is distinguishability. Although 1 is less than 2, we would not dare suggest any inequality in the two values (male, female) of the variable. Nor is there any meaning to the equal interval between 1 and 2 when applied to gender.

Levels of measurement are used to show the differences and extent to which numerical properties apply in objects that are measured. Different statistical analyses of numbers require, or assume, the presence of certain numerical properties. Otherwise, the analysis is inappropriate, invalid, and sometimes meaningless. The following levels of measurement are illustrated with examples.

Nominal. Data measured at the nominal level consist of named categories without any particular order to them. *Numbers are used here to name,* or distinguish the categories, and are purely arbitrary.

Variable: Political Party

Values: Republican: 1

Democrat: 2

Ordinal. Data measured on the ordinal level consist of discrete categories which have an order to them. No implication of equal intervals between numbers is made. Some variables do not admit of more precise measurement than simple ordering.

Variable: Pain

Values: None-O

Mild: + 1

Moderate: + 2

Severe: + 3

Excruciating: + 4

Continuous (Interval): Data measured at the continuous level can assume any value, rather than just whole numbers. In addition, we assume that equal, uniform intervals between numbers represent equal intervals in the variable being measured. An example is the Celsius scale of temperature, where the distance between 2 and 4 degrees is the same as the distance between 8 and 10 degrees.

Variable: Temperature

Values: The centigrade scale, originating at -273°

Continuous (Ratio): The mathematically strongest level is the ratio, where numbers represent equal intervals and start with zero. The existence of a zero value allows ratio statements, such as "4 is twice 2." Examples are provided by any absolute scale, or variable that has no negative numbers (less than zero) in its scale of measurement; these include height, weight, pH, blood pressure, and temperature measured on the Kelvin scale.

Variable: Temperature

Values: The Kelvin scale, originating at 0°

The order of the levels given is from the weakest to the strongest, where strength refers to the level of mathematical manipulation. A nominal level permits counting frequencies). An ordinal level allows use of the algebra of inequalities. Interval and ratio levels allow multiplication and division. Higher levels preserve and include the properties of lower levels. A variable that allows ratio measurement can always be reduced to a more primitive level of measurement, as seen with temperature, using an interval (Celsius) and ratio (Kelvin) level. We could also use an ordinal level with temperature, by devising the following scale:

0-Cold

l-Lukewarm

2-Warm

3-Hot

4-Extremely hot

We could even categorize temperature, and use numbers arbitrarily with no ranking indicated, for a nominal level:

Hot-l

Warm-2

Cold-3

Although higher levels of measurement can be reduced to lower levels, the reverse is not necessarily true unless the numerical properties of higher levels apply to the values of the variable. The variable *pain* cannot be measured above an ordinal level at this time. The lower, or more primitive, the level of measurement, the more restricted and less mathematically sophisticated is the statistical analysis.

Statistics and Parameters: The term *statistic* refers to a measure made on a sample, and is denoted by Roman letters, such as X or s. The term *parameter* refers to a measure made on a population, and is denoted by Greek letters, such as μ or σ. This distinction will be very important when discussing inferential statistics.

SIGNIFICANT FIGURES

By convention, the number of digits used to express a measured number is a rough indication of the error. For example, if a measurement is reported as being 35.2 cm, you would assume that the true length was between 34.15 and 35.24 cm (the error is about 0.05 cm). The last digit (2) in the reported measurement is uncertain, although we can reliably state that it is either a 1 or 2. The digit to the right of 2, however, can be any number (0,1,2,3,4,5,6,7,8,9). If the measurement is reported as 35.20 cm, it would indicate that the error is even less (0.005 cm). The number of reliably known digits in a measurement is referred to as the number of *significant figures.* Thus, the number 35.2 cm has three significant figures, and the number 35.20 has four. The numbers 35.2 cm and 0.352 m are the same quantities, both having three significant figures and expressing the same degree of accuracy. The use of significant figures to indicate the accuracy of a result is not as precise as giving the actual error, but it is sufficient for some purposes.

Zeros as Significant Figures

Final zeros to the right of the decimal point which are used to indicate accuracy are significant:

170.0 four significant figures

28.600 five significant figures

0.30 two significant figures

For numbers less than one, zeros between the decimal point and the first digit *are not* significant:

0.09 one significant figure

0.00010 two significant figures

Zeros between digits *are* significant:

10.5 three significant figures

0.8070 four significant figures

6000.01 six significant figures

If a number is written with no decimal point, the final zeros may or may not be significant. For instance, the distance between the earth and the sun might be written as 92,900,000 miles, although the accuracy may be only $\pm$ 5,000 miles. Only the first zero after the 9 significant. On the other hand, a value of 50 mL measured with a graduated cylinder would be expected to have two significant figures owing to the greater accuracy of the measurement device. To avoid ambiguity, numbers are often written as powers of 10 (using scientific notation) making all digits significant. Using this convention, 92,900,000 miles would be written 9.290×10^7, indicating that there are four significant figures.

Calculations Using Significant Figures

The least precise measurement used in a calculation determines the number of significant figures in the answer. Thus, 73.5 + 0.418 = 73.9 rather than 73.918, since the least precise number (73.5) is accurate to only one decimal place. Similarly, $0.394 - 0.3862 = 0.0078 \approx 0.008$ with only one significant digit, since the least precise number (0.394) is precise to only the nearest one thousandth (even though it has three significant figures).

For multiplication or division, the product or quotient has the same number of significant figures as the term with the fewest significant figures. As an example, in $28.08 \times 4.6/79.4 = 1.6268$, the term with the fewest significant figures is 4.6. Because this number has two significant figures, the result should be rounded off to 1.6.

ROUNDING OFF

The results of mathematical computations are often rounded off to specific numbers of significant figures. Rounding is done so that you do not infer accuracy in the result that was not present in the measurements. The following rules are universally accepted and will ensure bias-free reporting of results (the number of significant figures desired should be determined first).

- If the final digits of the number are 0, 1, 2, 3, or 4, the numbers are rounded down (dropped, and the preceding figure is retained unaltered). For example 3.51 is rounded to 3.5.
- If the final digits are 5, 6, 7, 8, or 9, the numbers are rounded up (dropped, and the preceding figure is increased by one). For example, 3.58 is rounded to 3.6.

DESCRIPTIVE STATISTICS

Although open to both misapplication and misinterpretation, statistics can provide us with the meaning in a data set. A data set is simply the list or group of numbers that results from measurement. Descriptive statistics offers a variety of methods for organizing data and reducing a large set of numbers to a few, informative numbers that will *describe* the data.

Data Representation

When a data set is obtained, it should be organized for inspection through use of a frequency distribution, which can represent the data both numerically and graphically. Regardless of how sophisticated the data analysis will be, taking a look at the data is one of the most useful procedures, and one of the simplest, to suggest further analysis and prevent inappropriate analysis.

In representing data, we distinguish a frequency distribution from a grouped frequency distribution. In a *frequency distribution,* the data are ordered, from the minimum to the maximum value, and the frequency of occurrence is given for each value of the variable. With a *grouped frequency distribution,*

values of the variable are grouped into classes. For example, if values range from 1 to 100, a frequency distribution that could be an ordered list of 100 different values is not practical or helpful. Usually 10 to 20 classes are a desirable number of classes.

Table 10-1 illustrates both an ungrouped and grouped frequency distribution. In constructing a frequency distribution, we first find the minimum and maximum values of the variable, list the values in order, tally the number of occurrences for each value in the ordered list, and calculate the percentage and cumulative percentages. The percentage is obtained from the frequency divided by the total number of values. The cumulative percentage accumulates the percentage of each value. For example, the value of 3 occurs three times, or 15%, ([3/20] x 100). The values of 2 and 3 account for 25% of the total observations in the distribution.

In the ungrouped frequency distribution, the class interval is actually one. Every value between the minimum and maximum is included. If the range of values is extremely large, group values into classes, and represent the frequency of each class, as shown in Table 10-1, for the same data set. The ability to "see" information in the numbers is lost with more than 20 intervals, and many prefer 10 to 12 intervals.

TABLE 10–1. FREQUENCY DISTRIBUTIONS, UNGROUPED AND GROUPED

Data Set 6, 5, 3, 7, 3, 2, 4, 6, 5, 4, 6, 4, 3, 2, 8, 7, 4, 5, 5, 5

Frequency Distribution:

Variable Values	*Tally*	*Frequency*	*Percentage*	*Cum Percentage*
2	11	2	10	10
3	111	3	15	25
4	1111	4	20	45
5	1111	5	25	70
6	111	3	15	85
7	11	2	10	95
8	1	1	5	100
	Total:	20	100%	

Grouped Frequency Distribution:

Interval	*Frequency*	*Percentage*	*Cum Percentage*
2–3	5	25	25
4–5	9	45	70
6–7	5	25	95
8–9	1	5	100
Total:	20	100%	

The goal of constructing a frequency distribution is to allow inspection of the data by summarizing, organizing, and simplifying the data *without misrepresenting the data.* A frequency distribution allows you to observe patterns or trends, and to begin extracting information about the numbers. In Table 10-1 we see that values of the variable tend to occur most frequently in the middle range, and to be relatively infrequent at extreme values (the “tails" of the distribution). The tally marks indicate this distribution. The percentage column shows the largest percentages for the middle values of 4 and 5. We also see that the cumulative percentage grows most rapidly in the middle range. As we add the value of 4, the cumulative percentage jumps from 25% to 45% , and adding 5 brings it to 70%. Almost three-fourths of the values are included at that point. A researcher would want to know how values distribute, and this may be of significance for interpreting results. If the frequency of values tended to be high at low values and more infrequent at high values, we would say that the distribution was skewed. If the variable was

test scores, then such a distribution tells us that the test was difficult, the students were poorly prepared, the students were lacking in ability, or all three!

If we wish to see, literally, how the data distribute, then graphic presentation of the frequency distribution is possible. The most basic forms are the histogram, the frequency and percentage polygon, and the cumulative percentage or ogive curve.

Histogram: This is a bar graph, where the height of the bar indicates the frequency of occurrence of a value or class of values.

Frequency Polygon: This is a graph in which a point indicates the frequency

of a value, and the points are connected to form a broken line (hence a polygon) .

*Percentage :*The numerical frequency on the Y-axis is replaced with the percentage of occurrence in this form of the polygon.

Cumulative Percentage Curve: This graph plots the cumulative percentage on the Y-axis against the values of the variable on the X-axis. The curve then describes the rate of accumulation for the values of the variable.

In using graphs, the horizontal axis is often referred to as the X-axis, and is the abscissa, whereas the vertical axis may be referred to as the Y-axis, or the ordinate.

Figure 10-1 illustrates a combined histogram and frequency polygon for the data set used in Table 10-1. Here the variable, X, is assumed to be continuous, so that the bars are joined at their bases, and the *real limits* of the variable values are indicated on the horizontal axis. The first bar is for the value 2, but if X is in fact a continuous variable, then X could have any value, so that 2 has a *width* on the real number line of 1.5 to 2.5. Alternatively, if X were a discrete variable, then the bars could be separated, and the numbers on the horizontal axis would most likely be whole numbers, or integers. If, instead of frequency we labeled the vertical axis with percentage, then the frequency polygon becomes a percentage polygon. Bar graphs are appropriate when representing the frequencies of the categories with a qualitative variable, or the frequencies of values with a quantitative but discrete variable. For example, the population of various countries is a qualitative variable whose frequencies, or size, could be indicated with a histogram, or a creative variation using different-sized figures called a *pictogram.* Likewise, the discrete variable, number of children per family, could be shown with a histogram.

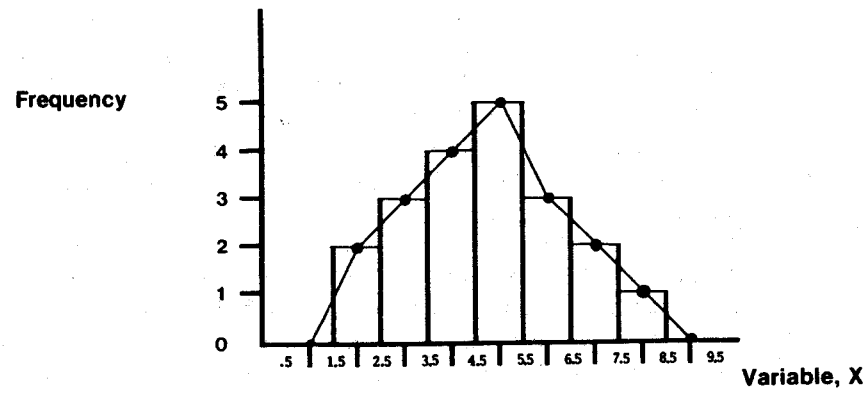


Figure 10-1. Histogram and frequency polygon for a continuous variable, X.

One good way to display a frequency distribution is with a pie chart (Figure 10-2). This way of representing the contribution of various parts to the whole was invented by Florence Nightengale, a pioneer in the fields of nursing and outcomes research.

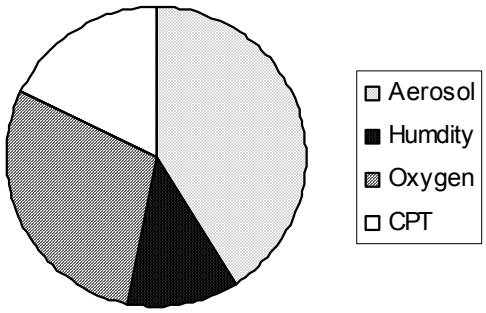


Figure 10-2. Frequency distribution displayed as a pie chart. Each area represents a percentage of the total. In this example, various therapies are shown as percentages of the total workload.

Figure 10-3 shows a percentiles plot. This is a graph showing the cumulative percentages of the data along the horizontal axis and the actual values of the data on the vertical axis. The value of a percentiles plot is that it summarizes complex raw data in a way that makes intuitive sense. For example, the top graph in Figure 10-3 shows the number of ventilators used daily for a number of months. But we cannot predict how many ventilators we need to own and how we may need to rent. The percentiles plot in the bottom graph shows that 85% of the time, we use 50 or fewer ventilators. Thus, if we purchase 50 ventilators we can expect to need rentals only 15% of the time.

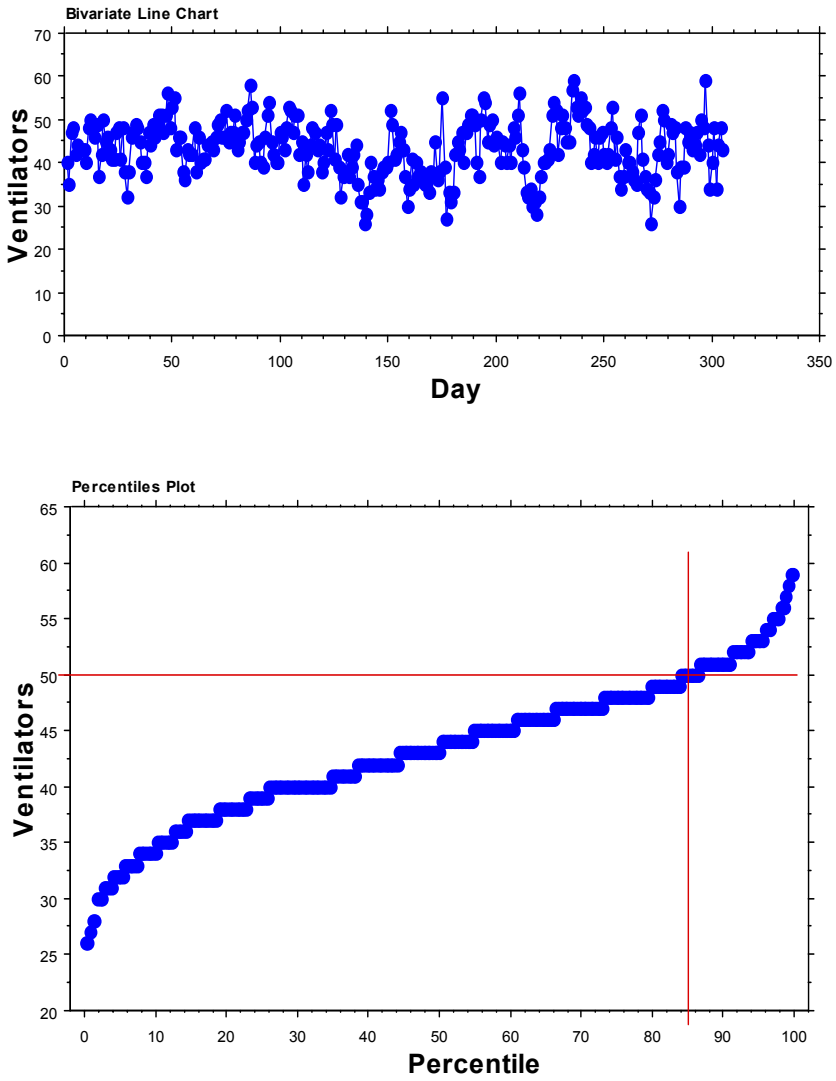


Figure 10-3. The top graph shows ventilator usage (number of ventilators per day on the vertical axis) over a range of several months (days on horizontal axis). The percentiles plot on the bottom shows the percentile (or percent of the days) that a given number of ventilators or less is used. The lines intersecting the plot show that 85% of the time, 50 or fewer ventilators are in use.

The graphic representation of a distribution of data is associated with a number of terms used to describe the *distributional form,* that is, the shape of the distribution. Thus, we can characterize an entire distribution with a single term. These terms are illustrated in Figure 10-4. The smooth curves in Figure 10-4 actually indicate that we have a continuous variable; every value can occur with some frequency along the horizontal axis, which gives the variable values. If we take the histogram in Figure 10-1 and make the class widths smaller and smaller, a smooth curve will ultimately result. Unless a variable is discrete or qualitative, we can represent the distribution of its values with such smooth, or continuous, curves. This is convenient for explanatory presentations of many statistical concepts, and will be used in the present chapter.

The form of the curve in Figure 10-4A is often referred to as *normal.* It is symmetrical and bell-shaped. However, it is only a normal curve if it represents a mathematical function known as the normal or Gaussian function, and this cannot be determined from inspection. A rectangular shape as in Figure 10-4B results from a uniformly distributed variable such as the frequency of values in a deck of cards. Skewness is seen in parts C and D, and indicates that values of the variable occur more frequently at either end of a distribution. A bimodal distribution, as in E, occurs when there is no single most-frequent value. For instance, if there is a high and low cluster of IQ values in a large group, we could see a bimodal, or more generally, a multimodal distribution. In F, kurtosis is represented; this term describes the peakedness or flatness of the distribution.

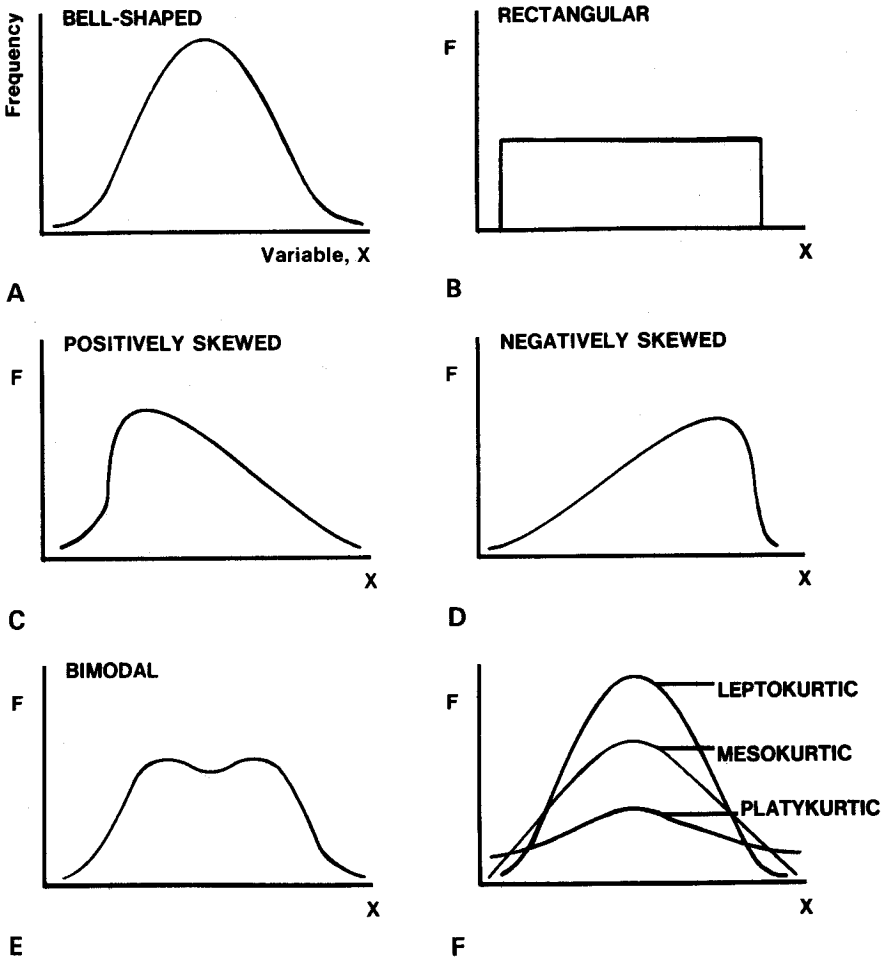


Figure 10-4. Illustration of various distribution shapes.

Measures of the Typical Value of a Set of Numbers

Three statistics are used to represent the typical value (also called the central tendency) in a distribution of values. Each statistic is a single number or an index that characterizes the center, or average value, of the whole set of data. These statistics are the mean, the median, and the mode.

The Summation Operator. Summation is a mathematical operation common in statistical calculations. The summation operator is denoted by the Greek capital letter, sigma (Σ), and simply indicates addition over values of a variable. The general representation of the operator is

$$\sum_{i=1}^{n} X_i$$

which is read "the summation of X_i for i equal 1 to n". Summation simply means, add the values of X_1 through X_n:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \ldots + X_n$$

The subscripted variable, X_i which is read "X sub i," is used to denote specific values of the variable. For instance X_1 is one value of X_i, X_2 another, and so on. An example should make the use of the summation operator clear. Let the variable X have three values, each of which is given by a subscripted variable: $X_1 = 3$, $X_2 = 4$, $X_3 = 2$. Then,

$$\sum_{i=1}^{3} X_i = X_1 + X_2 + X_3 = 3 + 4 + 2 = 9$$

The Mean. The mean is the sum of all the observations divided by the number of observations. The symbol for the mean is $\overline{X}$ or μ, depending on whether we have a sample (statistic) or population (parameter) value respectively. The formula for the mean represents the definition in briefer form:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

were X is the variable and there are n values. For example, let us use the above values of X (3, 4, and 6) so that $n = 3$. Then the mean is:

$$\overline{X} = \frac{(3 + 4 + 2)}{3} = 3.0$$

The Median. The median is the 50th percentile of a distribution, or the point that divides the distribution into equal halves. The median is the value below which 50% of the observations occur. For grouped observations, a formula is used to find the exact value of the median. For an *odd* number of observations, the median is the value that is equal to $(n + 1)/2$, where n is the number of observations. For example, if we have 1, 3, 5, 6, and 7 as our data, n is 5. The median is $(5 + 1)/2 = 5$, the third value. Notice this formula gives the number of the observation, not its value.

With an *even* number of observations, the median is equal to the sum of two middle values divided by 2. For example, if we have 1, 3, 5, 6, 7, and 9 as data, the median is $(5 + 6)/2 = 5.5$. This formula gives the value of the observation that divides the data set.

The Mode. The mode is the most frequently occurring observation in the dis- tribution. The mode is found by inspection, or counting. In Table 10-1 the mode is the value 5, which occurs 5 times. In a histogram, the mode is the highest bar.

Interpretation and Application. The mean is the most sophisticated measure of central tendency, is influenced by every value in the distribution, and assumes at least an interval level of measurement, to perform addition and division. The mean is inappropriate for a qualitative variable with a nominal or ordinal level of measurement. What sense is there in calculating the mean of political affiliations, which have the values 1, 2, and 3 for Republican, Democrat, and other? Of course, we could calculate a number, but it is inappropriate for the variable. The mean is termed an *interval* statistic.

The median is considered an *ordinal* statistic, and is less sophisticated than the mean. The median requires only ranking of numbers, not equal intervals. If a single value skews the value of the mean, the median may give a more typical representation of the data than the mean. For instance, if salaries range from $10,000 to $14,000, and one person earns $25,000, then the mean will be in- creased, or skewed, while the median will probably better represent the average salary.

The mode is a *nominal* statistic, since it requires only a nominal level of measurement. The qualitative variable, political affiliation, is a good example for using the mode to typify the observations. Which category occurs most frequently with the qualitative variable? With only a nominal level, we cannot (at least appropriately) calculate a median or a mean.

Measures of Dispersion

Most research studies provide at least two descriptive statistics: one measure of central tendency and one measure of dispersion. Measures of dispersion indicate the variability, or how spread out, the data are and include the range, variance, standard deviation, and coefficient of variation. The need for a measure of central tendency *and* dispersion to characterize more fully a distribution is seen in Figure 10-5. We have two different distributions of pH values, both with the same mean. Although both center on the same value, they do not "distribute" the same. The range and variability are different. If we had only the mean (7.40) to characterize the data, we would conclude the distributions are the same. But while they are the same with regard to their central tendency, they are quite different in their dispersion. Note that pH is a continuous variable and measurable at a ratio level (there are a true zero and equal intervals).

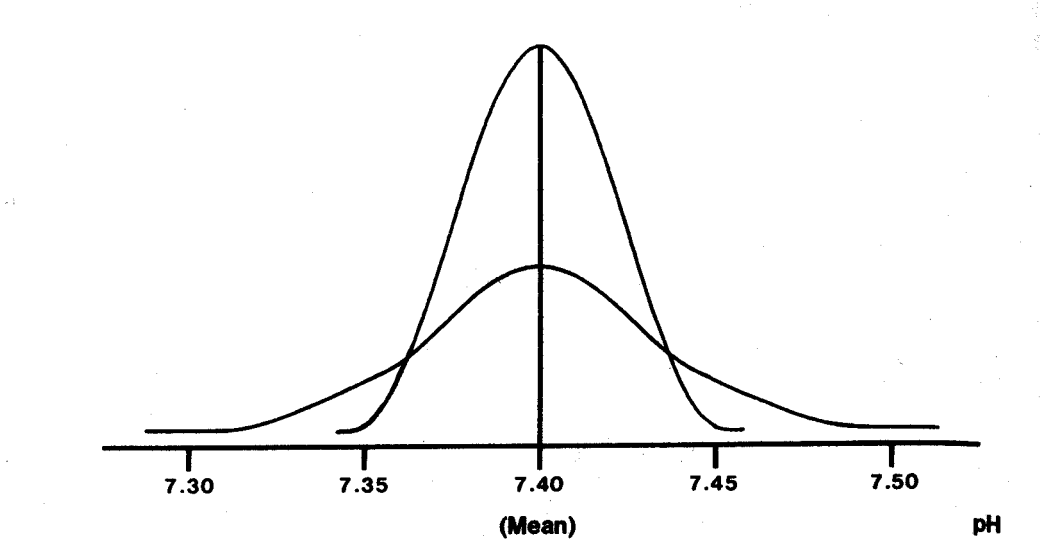


Figure 10-5. Two distributions of pH values, with the same mean but different amounts of dispersion.

Range. The range is the distance between the smallest and the largest values of the variable. Range = X_{max} -X_{min}, where X_{max} and X_{min}. are the maximum and minimum values in the distribution. The range is the simplest measure of dispersion, and is very informative to a researcher. Although the range is the actual distance between the smallest and largest values, the actual minimum and maximum values themselves are usually more informative.

Variance and Standard Deviation. The variance is a measure of how different the values in a set of numbers are from each other. It is calculated as the average squared deviation of the values from the mean. The standard deviation is the square root of the variance. The standard deviation has the same units as the original measurements while the variance does not. Equations and an example of the variance and standard deviation are given in Table 10-2. There is a difference in the equation for the variance, depending on whether we have a sample with *n* observations, or a population with the total collection of *N* observations.

TABLE 10–2. THE VARIANCE, STANDARD DEVIATION, AND COEFFICIENT OF VARIATION WITH A SET OF SAMPLE VALUES

Data Set 2, 3, 5, 6

Variance

$$\text{Variance, } S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}$$

Calculation:

$\bar{X} = \Sigma X_i / n = 16/4 = 4.0$

Value of **X**	$(X - \bar{X})$	$(X - \bar{X})^2$
2	−2	4
3	−1	1
5	1	1
6	2	4
	$\Sigma = 0$	$\Sigma = 10$

Sample variance, $S^2 = 10/(4 - 1) = 3.33$

Standard Deviation

$$\text{Standard deviation, } S = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}} = \sqrt{S^2}$$

Sample standard deviation, $S = \sqrt{3.33} = 1.83$

Coefficient of Variation

Coefficient of Variation, C.V. = $(S/\bar{X}) \cdot 100$

C.V. = $(1.83/4.0) \cdot 100 = 45.6\%$

For a *sample,*

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}$$

For the *population,*

$$\sigma^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}$$

We use Roman letters, S^2 and $\overline{X}$ for the variance and mean with a *sample,* but Greek letters σ^2 and μ to indicate the *population* variance and mean. The use of (n –1) when calculating sample variance is needed to obtain an unbiased estimate of the true population variance from the sample values. With a large *n,* such as 100, the numerical difference between n and n -1 tends to diminish, and use of n -1 does not greatly affect the calculation of the variance. Further explanation of biased estimators and the reason for using n -1 can be found in statistical textbooks. The calculations in Table 10-2 assume sample values. We first calculate the mean, and then take the difference or deviation for each X_i for i from 1 to n, square the deviations so all are positive, and then sum the squared deviations. After dividing by (n – 1), we have the variance, which is simply the *average squared deviation* from the mean. Notice the sum of the second column ($X - \overline{X}$) should always add to zero, since the mean is the center of a distribution and differences above and below the mean will then cancel.

As a squared value, the variance is always a positive number. By taking the square root, we have the standard deviation, which is then the *average deviation* from the mean. The larger the variance and standard deviation are, the more dispersed are our values. For example, In Figure 10-5 the narrower distribution of pH values might have a standard deviation of 0.02, while the wider distribution may be 0.05.

When testing hypotheses and estimating sample sizes, it is often necessary to calculate a *pooled standard deviation* (S_p):

$$S_p = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$$

where

S_p = the pooled standard deviation

S_1^2 = the variance of the first sample

S_2^2 = the variance of the second sample

n_1 and n_2 are the two sample sizes.

Coefficient of Variation. The coefficient of variation expresses the standard deviation as a percentage of the mean. The equation for the coefficient of variation is given in Table 10-2, and is simply the sample standard deviation divided by the sample mean. The coefficient of variation is not useful as a single value, but is applicable when we wish to *compare* the dispersion of observations, for example, with two different instruments or methods intended to measure the same variable. This statistic can also be helpful in comparing the dispersion of values at high and low values of a variable. For example, the dispersion of measured values is large at low creatinine levels, but shrinks at higher creatinine levels. The data suggest that an instrument to measure creatinine has more random error (dispersion is greater) at the low range of creatinine values.

Standard Scores. A standard score, or z score, is a deviation from the mean expressed in units of standard deviations.

$$z = \frac{X - \overline{X}}{S}$$

where

z = the z score

X = an individual value from the data set

$\overline{X}$ = the mean value of the data set

S = the standard deviation of the data set

A z score of +20 means that the value of X is two standard deviations above the mean. A z score of +1.6 is likewise 1.6 standard deviations above the mean. Standard scores offer a way to express raw values in terms of their distance from the mean value, using standard deviation units.

The data in Table 10-2 have a mean value of 4.0. Calculating the z or standard score for the raw value of 2 is as follows:

$$z = \frac{2 - 4.0}{1.83} = -1.09$$

This value of z indicates that the value of 2 is 1.09 standard deviations *below* (indicated by the negative sign) the mean. Standard scores will be useful when obtaining percentages and probabilities with certain well-known distributions such as the normal.

Correlation and Regression

A coefficient of correlation is a descriptive measure of the degree of relationship or association between two variables. This is the first statistic that involves *two* variables. The concept of correlation implies that two variables co-vary. That is, a change in variable X is associated with a change in variable Y. The most common correlation coefficient with a continuous variable measurable on an interval level is the Pearson product-moment correlation coefficient (the Pearson r). Another basic assumption of the Pearson r is linearity: The relation of the two variables is linear. Visual inspection of a plot of coordinate points for the X and Y variables is necessary to confirm linearity or to determine non-linearity. Such a plot is called a *scattergram*, and is illustrated in Figure 10-6 for both a linear and a curvilinear relationship between two variables.

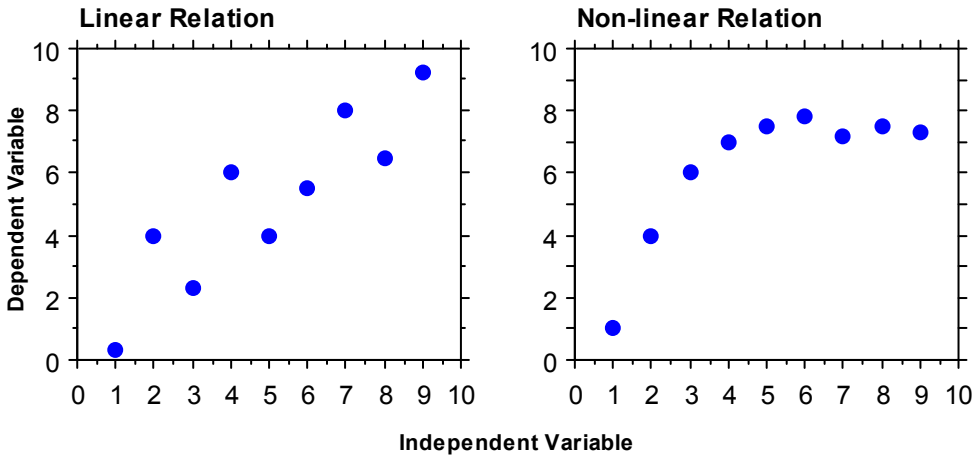


Figure 10-6. Illustration of scattergrams for both linear (left) and curvilinear (right) relations between the variables *X* and *Y*.

In Figure 10-6B, the Pearson *r* value would erroneously underestimate the degree of relation between *X* and *Y* because of its nonlinear nature. Calculation of the *r* value without inspection of the scattergram in Figure 10-6B could have led you to conclude that *X* and *Y* have a weak relationship, or none at all, when in truth they are clearly related.

The Pearson *r* statistic ranges in value from –1.0 through 0, to +1.0 and indicates two aspects of a correlation, the *magnitude* and the *direction.* Figure 10-7 illustrates some possible values for a Pearson *r* and their meanings. A positive value indicates a positive or direct relation *Y* increases as *X* increases (Figure 10-7A). A negative value for *r* indicates an inverse relation: *Y* decreases as *X* increases. (Figure10-7B.) The closer the absolute value is to 1.0, the stronger and more perfect the relation, while a value approaching zero indicates a lack of relationship, as in Figure 10-7D. In Figure 10-7D, low values of *X* are seen to correspond to high and low values of *Y*. The same is true for high values of *X*. Thus there is no systematic co-varying suggesting *X* and *Y* are related.

As a rule of thumb, the correlation between *X* and *Y* is considered *weak* if the absolute value of *r* is between 0 and 0.5, *moderate* between 0.5 and 0.8, and *strong* if between 0.8 and 1.0.

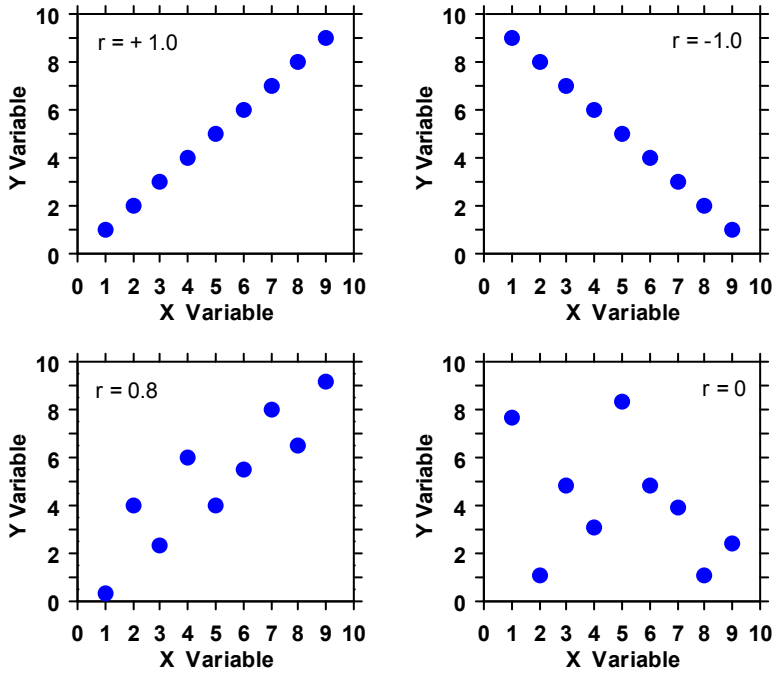


Figure 10-7. Illustration of four Pearson *r* values, indicating varying strengths of relation between variables *X* and *Y*.

Table 10-3 gives an example of the calculation of *r*, although nobody calculates the statistic by hand these days. Even the simple statistics like mean and standard deviation are usually calculated with computer spreadsheets of specialized statistical software.

TABLE 10–3. CALCULATION OF THE PEARSON PRODUCT–MOMENT CORRELATION COEFFICIENT

$$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \, \Sigma y^2}}, \quad x = X - \overline{X}, \quad y = Y - \overline{Y}$$

X	Y	$(X - \overline{X})$	$(Y - \overline{Y})$	x^2	y^2	xy
1	2	−3	−2.2	9	4.84	6.6
3	3	−1	−1.2	1	1.44	1.2
2	4	−2	−.2	4	.04	.4
6	5	2	.8	4	.64	1.6
8	7	4	2.8	16	7.84	11.2
$\overline{X} = 4.0$	$\overline{Y} = 4.2$			$\Sigma x^2 = 34$	$\Sigma y^2 = 14.8$	$\Sigma xy = 21.0$

$$r = \frac{21.0}{\sqrt{34 \cdot 14.8}} = .936$$

When there is a linear relationship between two variables, we often wish to use the value of one variable (which may be easy to measure) to predict the value of the other variable (which we cannot easily measure). Of course, both variables had to be measured at some time to establish the presence of a correlation and prediction equation. The procedure is called least squares, simple regression analysis. When we measure *X* and predict *Y*, *Y* is said to be *regressed* on *X*. The term *simple* indicates that *Y* is predicted from only one variable and not several simultaneously, which would involve multiple linear regression. Essentially, a line of best fit is that with the minimum distance to all of the data points. This is referred to as a *least squares* criteria for fitting the line because it is the line that minimizes the squared distance between points on the line and the actual data points on the graph. Figure 10-8 illustrates this line for the correlation data in Table 10-3. Such a line can be describes with a linear equation of the form

$$\hat{Y} = a + bX$$

where $\hat{Y}$ is the *predicted* value of *Y* for the given value of *X*. The letter *a* stands for the *Y*-intercept (the value of Y when X equals zero). The letter *b* stands for the slope of the line (the change in *Y* for a given change in *X* or $\Delta Y/\Delta X$).

Computer programs often give a value for r^2 along with the regression equation as a measure of how well the line fits the data. The r^2 statistic is called the *coefficent of determination.* The value of r^2 ranges from 0 to 1.0 and interpreted as the proportion of the variation in *Y* that is explained by the variation in *X*. To understand what this means, consider that the total difference in *Y* values relative to their mean value has two components. One component is due to the linear relationship between *Y* and *X*. For example, as the independent variable *X* increases, the dependent variable *Y* may also increase as predicted by the regression equation. If a perfect correlation existed between the two (ie, $r = 1.0$ and all of the *X,Y* points lie on the regression line) then *all* of the variation in *Y* would be explained by the variation in *X* (that is, $r^2 = 1.0$). We say that the difference between an individual value of *Y* and the mean value for *Y* is "explained" by the variation in *X*.

However, when less than a perfect correlation exists, some of the measured *Y* values will lie away from the regression line. This second component in the difference between an individual value of *Y* and the mean value of *Y* is the difference between the measured value of *Y* and its corresponding predicted

value, $\hat{Y}$. This difference is "unexplained" by the variation in X values. The value of r^2 can be thought of as:

$$r^2 = \text{explained difference/total difference}$$

Thus, the worse the correlation, the more the unexplained difference is relative to the total difference, making the explained difference smaller, and the value for r^2 smaller.

Measures of correlation exist for lower levels of measurement. The Spearman rank coefficient can be used with ordinal levels of measurement, and the phi coefficient with nominal levels.

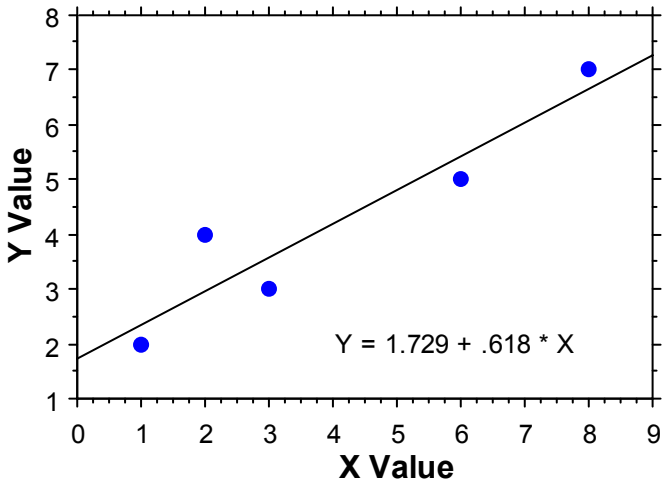


Figure 10-8. An example of simple linear regression, giving the line of best fit for data in Table 10-3.

INFERENTIAL STATISTICS

Although a sample from a population is economical, we still wish to use the sample measurements (statistics) to *infer* to the population measures (parameters). Inferential statistics offer methods for inference by combining probability with descriptive statistics. Essentially, inferential statistics allow us to make probability statements about unknown population parameters based on the sample statistics. A basic understanding of probability forms the basis for an explanation of inferential statistics.

The Concept of Probability

The probability of an event can be defined as the relative frequency, or proportion, of occurrence of that event out of some total number of events. Since probability is a proportion, it always has values between 0 and 1, inclusively. For example, the probability of obtaining an ace from a deck of well-shuffled cards is 4/52, or 0.077. There are 4 aces (the event) out of a total of 52 cards, or events.

When a frequency distribution gives the relative frequencies of each value of a variable, it is actually a *probability distribution.* The concept of a probability distribution is essential to inferential statistics, and can be easily understood by use of a discrete variable.

The distribution of values for a discrete variable is called a *discrete distribution.* Let the discrete variable be the number of heads in five flips of a fair coin. Values of this variable can range from 0 heads to 5 heads. We then perform the sequence of five flips 32 different times, and obtain the frequency

distribution in Table 10-4. Since probability is simply relative frequency of occurrence, we can obtain the probability of each value of the variable (number of heads in five flips) as shown in column 3. For instance, the probability of the variable *X* having the value 0 is l out of 32, or 0.03125.

TABLE 10–4. PROBABILITY DISTRIBUTION AND CUMULATIVE PROBABILITY DISTRIBUTION FOR THE NUMBER OF HEADS IN FIVE FLIPS OF A COIN

X, No. of Heads	Frequency	$P(X = x)$	$P(X \leq x)$
0	1	1/32 (.03125)	1/32 (.03125)
1	5	5/32 (.15625)	6/32 (.1875)
2	10	10/32 (.3125)	16/32 (.5)
3	10	10/32 (.3125)	26/32 (.8125)
4	5	5/32 (.15625)	31/32 (.96875)
5	1	1/32 (.03125)	32/32 (1.0)
	32	32/32	

Several helpful mathematical conventions are shown in Table 10-4. First, the upper-case letter *X* denotes a *variable,* while the lower-case letter *x* denotes the *value* of the variable. Second, the symbol *p* denotes probability. The expression $P(X = x)$ is read "the probability that the variable *X* has the value *x."* Likewise, $P(X \leq x)$ is read "the probability that the variable X is less than or equal to the value *x."*

Column 3 gives probabilities of exact values of X. Column 4 provides *cumulative* probabilities. For instance, the probability that X is less than or equal to 2, $P(X \leq 2)$, is the sum of the probabilities that X is equal to 0, X is equal to 1, and X is equal to 2. The sum of these probabilities is: 1/32 + 5/32 + 10/32 = 16/32, or 0.50. In other words half of the values are between 0 and 2.

If we graph the relative frequencies in Table 10-4, we can obtain a visual representation of the probability distribution (Figure 10-9). A bar graph is used because we have a discrete variable. The height of a bar indicates how probable or improbable a value is. The values 0 and 5 are rare, or improbable, because they occur infrequently. The most probable events are 2 and 3. Just as we obtained the cumulative probabilities in the table, we can *sum* the probabilities represented by the height of each bar in Figure 10-9. This is in effect adding up the bars in Figure 10-9. We can obtain cumulative probabilities by summation when we have a discrete variable, but we must use the integration of calculus to "add up" the probabilities when we have a continuous variable. This difference, as well as the similarity, is seen in Figure 10-10. In the continuous case, the probability that *X* is less than or equal to 2 is given by the *area under the curve* between 0 and 2. This area is obtained by integration, not summation. The total area under the curve is 1.0, which indicates that there is a 100 percent likelihood that *X* will have a value between 0 and 5, which it must! Throughout the discussion on inferential statistics, we will use continuous distributions, as in Figure 10-10B, to present or illustrate probabilities. For a continuous probability distribution, the probability of any particular value is zero and the probability of an interval does not depend on whether or not either of its endpoints are included. For example, in Figure 10-10B, $P(X = 2) = 0$ and $P(X \leq 2) = P(X > 2)$.

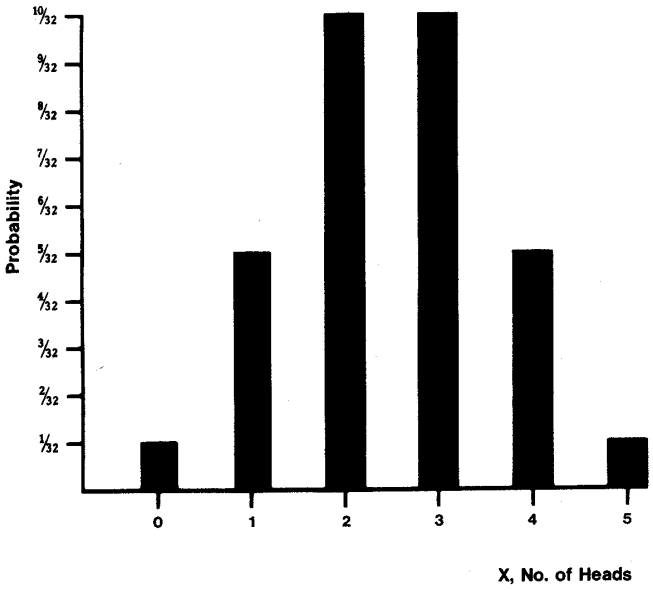


Figure 10-9. Discrete probability distribution for the number of heads in five tosses of a coin.

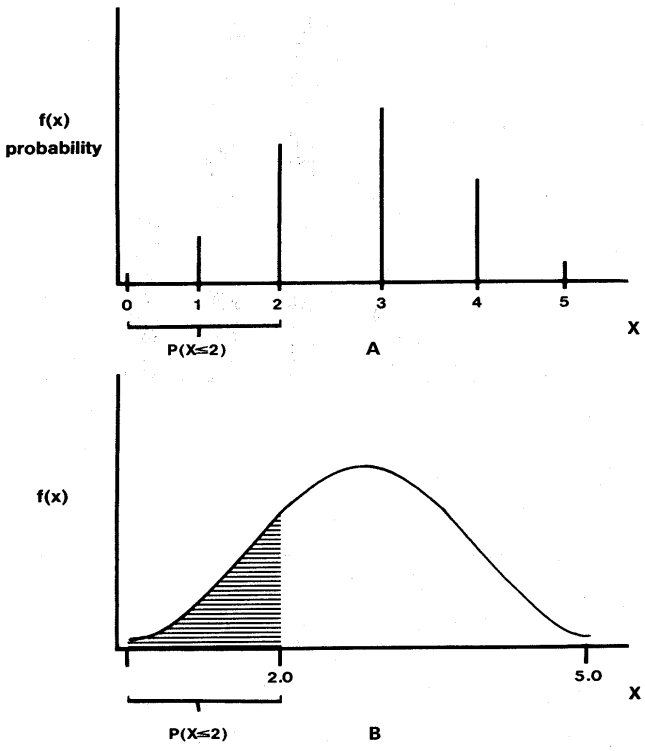


Figure 10-10. Comparison of cumulative probabilities for a discrete distribution (A) and for a continuous distribution (B). The Symbol f(x) means that the values on the vertical axis (probability in these graphs) are a function of the horizontal values of x.

The key to obtaining probabilities is to have the probability distribution. Many biological variables such as height, weight, or pH follow a distribution known as the normal distribution. This is a particular distribution described with a certain mathematical function (the Gaussian function), and from which we can obtain probabilities whenever we are willing to assume a variable follows this distribution. The normal is a continuous, not a discrete distribution. Other probability distributions will also be used when discussing inferential statistics. First we illustrate the use of the normal distribution for determining probabilities.

The Normal Distribution and Standard Scores

Earlier we defined standard or z scores, and now we will use such scores with the normal distribution. Figure 10-11 illustrates the areas under the normal curve. In a normally distributed variable, the mean is at the center of the distribution, and therefore, the mean is also the median and the mode.

The curve in Figure 10-11 is called a standard normal curve because the areas are indicated for standard deviation units from the mean. Thus, the mean itself, regardless of its value from actual data, is zero standard deviations away from itself. The z score for the mean must always be zero. In other words, points on the curve are actually z scores. Rather than recalculate areas under the curve for particular values of the mean and standard deviation, we use a general form of the curve with standard deviation units.

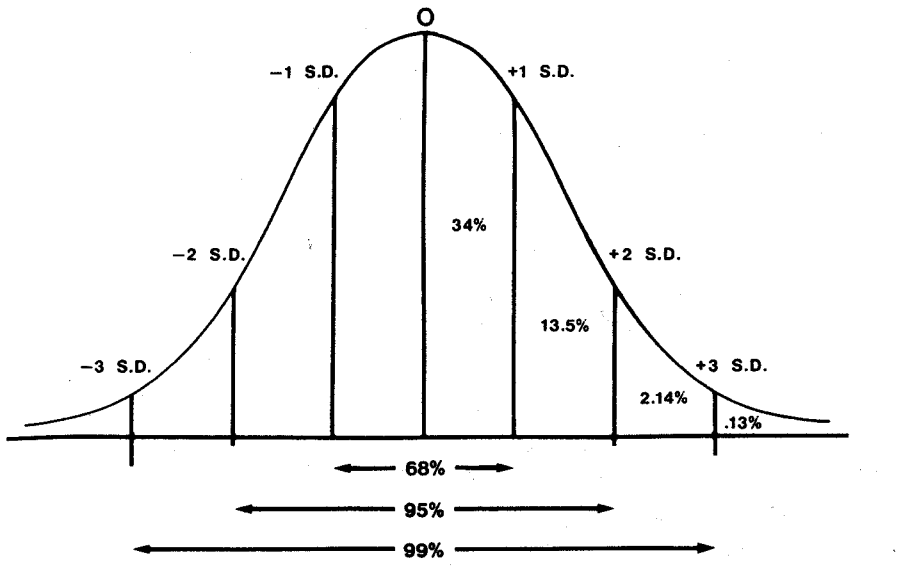


Figure 10-II. Approximate areas under the normal curve within one, two, and three standard deviations around the mean.

In a normal curve, approximately 68% of all the values are included in the area around the mean spanning $\pm$ 1 standard deviations. The area encompassed by $\pm$ 2 standard deviations from the mean corresponds to approximately 95% of the values in the distribution. The interval of $\pm$ 3 standard

deviations encompasses 99.7% of all the values. Exact areas, which are the probabilities for particular z score values, can be obtained from a table in a statistics book. or from a computer program.

An application of z scores and the normal distribution will identify the usefulness of knowing areas under the normal curve. We will use the continuous variable, pH, and assume its values follow a normal distribution. Further, pH in humans has a population mean of 7.40 and a standard deviation of 0.02. The distribution of pH values is seen in Figure 10-12. Both original units and z scores are given. Since pH is normally distributed, we can say that approximately 95% of the population's values will be between 7.36 and 7.44, because that is a width of $\pm$ 2 standard deviations. We could also say that only approximately 2.5% of the population will have a pH greater than 7.44, or the probability that a pH value is above 7.44, *P(*pH > 7.44), is less than 0.025. The percentile statement tells what percentage of observations are less than the value of 7.44, and the probability statement gives the proportion of observations above or below a point, depending on the direction of the statement. Notice that we can determine from the distribution which values or range of values are likely, and which unlikely. If 95% of the values are between 7.36 and 7.44, then values *outside* this range, such as 7.33 or 7.46, are *not* likely. This reasoning is the basis for establishing normal ranges of many clinical variables. An appropriate table or computer gives exact areas under the normal curve for z scores within 3 standard deviations of the mean.

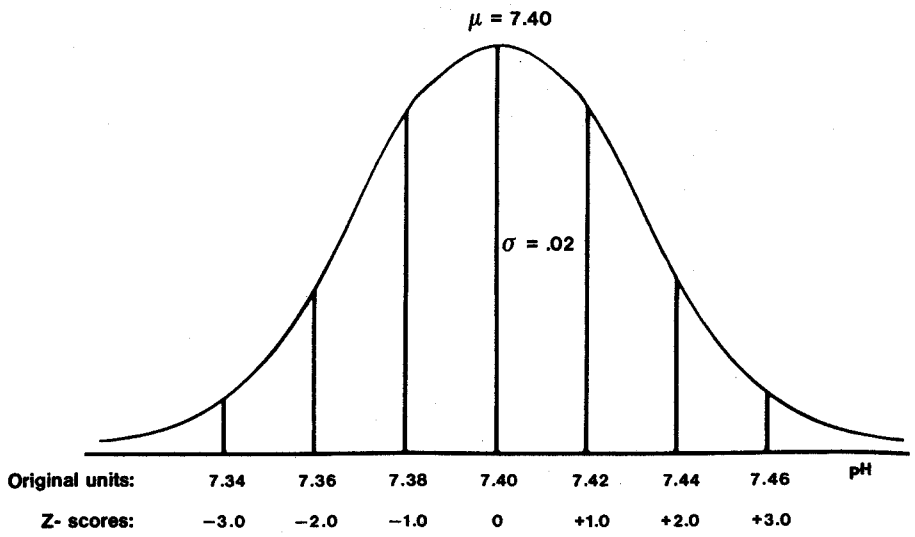


Figure 10-12. Normal distribution of pH values, with a mean of 7.40 and a standard deviation of 0.02.

For example, what is the probability of obtaining a pH value greater than 7.43? First, we need the z score for the value 7.43, that is, z = (7.43 - 7.40)/0.02 = + 1.5. The value 7.43 then is 1.5 standard deviations above the mean. Now we ask what is the probability of a z score greater than 1.5, *P(*pH > 7.43) = *P(z* > 1.5). The nor- mal curve is symmetrical around the mean, which is a z score of 0. A z score of 0, for example, has a value of 0.50, indicating half of the total curve area is accumulated at this midpoint.

For a probability of a z score greater than 1.5, we could use a statistical table. Alternatively, we could use Microsoft Excel to do a simple spreadsheet calculation using the cell equation:

=NORMSDIST(z)

where z is the desired number of standard deviations. If we substitute 1.5 for z, the cell evaluates to 0.9332, which indicates that 93.32% of the total area under the normal curve is found between negative infinity and 1.5 standard deviations above the mean. The area remaining under the curve *above* a z score of + 1.5 will be equal to 1.0 - 0.9332 = 0.0668. These areas corresponding to a z score of + 1.5 are illustrated in Figure 10-13. The probability of a pH value above 7.43 is 0.0668. Alternatively, we could say that approximately 93% of the population has a pH *less* than 7.43, and 7% above 7.43; or a pH of 7.43 is in the 93rd percentile.

To summarize, if a variable is normally distributed, then we can standardize any value of the variable by calculating a z score. Then we find the probability of a value greater, or less than, the given value from a table or computer program. *Remember that these probabilities will be accurate only if the variable does in fact have a normal distribution.*

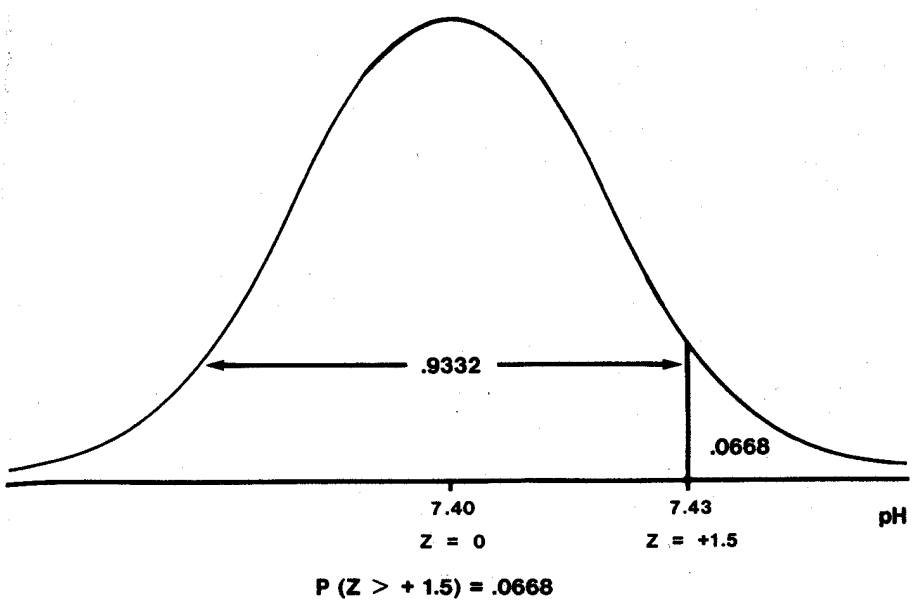


Figure 10-13. Areas under the normal curve for a pH value of 7.43 with a corresponding z score of +1.5.

Sampling Distributions

Since we defined probability as the relative frequency of an event, we were able to use the frequency distribution of a variable as a probability distribution. We have used a particular probability distribution known as the normal, or Gaussian, distribution. The single most important concept in inferential statistics is that of a sampling distribution of a statistic.

Definition. A sampling distribution of a statistic is the distribution of all values of that statistic when it is computed from random samples of the same size from a population.

A sampling distribution is the probability distribution of a *statistic.* We assumed that the variable pH has a normal distribution, and then we were able to calculate probabilities from this distribution. A statistic such as the mean, $\overline{X}$, which is calculated from a sample, will also have different possible values in different samples when the samples are randomly drawn from a population. These different possible

values of $\overline{X}$ will also have some distribution, called a sampling distribution. The mean, $\overline{X}$, can have a normal or a *t* distribution, depending on certain conditions which we will present.

An example of a sampling distribution may clarify this essential. Suppose that we have the continuous variable, National Board Scores, and the distribution of values for the *population* is nor- mal, with a mean, μ, of 100, and a standard deviation, σ, of 15. We have already seen that we can obtain probabilities for particular values of the variable. Now let us take a random sample of size 9, and calculate the sample mean, $\overline{X}$. Then we take another random sample of nine scores, and calculate a second $\overline{X}$, and repeat this procedure several thousand times. We would have a *distribution* of different $\overline{X}$ values.

The values are different because of *sampling error,* which is a random error. Even though the population has a mean of 100, our sample means are going to differ more or less from 100 due to random chance. The nine particular scores in a given sample are not likely to have a mean of *exactly* 100. However, the distribution of the sample means does center on 100, which is the original population mean. The mean of the sample means (symbolized as $\mu_{\overline{X}}$) is also 100. Where the population distribution has a standard deviation to describe its variability, the sampling distribution of the mean has a *standard error of the mean* (SEM), which is symbolized by $\sigma_{\overline{X}}$:

$$SEM = \sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and *n* is the sample size. If (as is usually the case) σ is unknown, *S* is substituted in the above equation. The SEM is simply the standard deviation of the values of $\overline{X}$.

We would expect the dispersion of sample means to be less than the dispersion of population values, and indeed, the equation shows that the SEM is smaller than the population standard deviation, σ. In fact, as the sample size increases, the smaller the SEM becomes. Finally, we can convert any specific value of the mean into a standard or z score with

$$z_{\overline{X}} = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$

We divide the difference between the sample mean and the population mean by the standard error of the mean. This quotient gives us the number of standard errors the sample mean is above or below the population mean. Just as with z scores for values of a variable, so we can use this z score for the mean with the normal curve to determine probabilities. For instance, if the population mean is 100 with a standard deviation of 15, what is the probability of obtaining a sample mean of 110 or greater for a sample size of 9? The steps to obtain a probability from the sampling distribution are the same as before:

1. Convert the sample mean value to a z score.
2. Find the probability for the z score from a table or computer program.

The z score for a mean of 110 is

$$z_{\bar{X}} = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{110 - 100}{15 / \sqrt{9}} = +2.0$$

Again using Microsoft Excel we fill a cell with the equation

=NORMSDIST(2.0)

When we hit the ENTER key, the cell shows the value 0.9772. This means that the area under the normal curve up to a value of 2 standard errors above the mean represents 97.72%. The area remaining in the upper tail of the distribution is 1.0 – 0.9772 = 0.0228. Thus, the probability of a sample mean of 110.0 or greater is the same as the probability of a z score of 2.0 or greater, which is 0.0228 or 2.28%. We would conclude that a value of 110.0 is not very likely to occur.

Inferential statistics use sampling distributions, which are probability distributions of a statistic, to make conclusions about *population parameters* based on *sample statistics.* Obviously, if we measure an entire population, we have a parameter, and would not need inferential statistics to make a probability statement about that value.

The t Distribution: The use of z scores and the normal distribution are usually based on the assumption that samples sizes are "large" and that we know the sample variance. Large usually means more than 30. However, a large sample size is frequently difficult to obtain in medical research. Therefore, we need another distribution, the *t* distribution. It looks like the normal distribution but more spread out. In fact, there is a whole family of *t* distributions whose shape depends on the "degrees of freedom" (defined for this purpose as one less than the sample size; $n - 1$). The smaller the sample size, the more spread out the distribution looks. However, that will not concern us unless we have to look up values in statistical tables. From here on out, we will assume that all statistical values will be calculate for us using a computer, either Microsoft Excel or some specialized statistical software package. We will generally use the *t* distribution rather than the normal. But the same ideas we discussed above apply because as the sample size gets larger and larger, the *t* distribution turns into the normal distribution. The *t* statistic is calculated like the z statistic except that we use the sample standard deviation:

$$t = \frac{\bar{X} - \mu}{S / \sqrt{n}}$$

where

$\bar{X}$ = the sample mean,

μ = the assumed population mean

S = the sample standard deviation

n = the sample size

In the situation where you are not comparing a sample mean with a population mean, but are comparing two sample means, the *t* statistic is calculated as:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$\overline{X}_1$ = the mean of one sample and

$\overline{X}_2$ = the mean of the other sample

S_p = the pooled standard deviation

n_1 and n_2 are the two sample sizes

Confidence Intervals

The mean and standard deviation of a sample are called *point estimates*, which are just single-value guesses about the population parameters. As we saw above, repeated calculations of a statistic like the sample mean result in a distribution of values. Somewhere in that distribution will be the true value of the population mean. The problem is, we have only performed one experiment and calculated one sample mean. We don't know if the true population mean is above or below our sample mean, nor how far away it is (how much error is in our estimate).

We accept that point estimates have built-in error, so we need to decide how much confidence to place in them. In medicine the convention is to keep the error less than or equal to 5%. That means we would like a confidence level of 95%.

Thinking in terms of the sampling distribution, a confidence level of 95% means that we expect the true population parameter (the mean in this case) to be in an area that makes up 95% of the area under the curve. Since we do not know if our sample mean is above or below the true value, we assume that there is equal probability that it lies below it as above it. In other words, we assume that our mean value lies within a certain number of standard errors above or below the true value. If we were using the normal distribution, it would be two standard errors above and below the mean. But for the *t* distribution, the number of standard errors depends on the sample size. For now, lets just say the sample mean lies within *t* standard errors of the true value with 95% confidence.

If our sample mean is within *t* standard errors from the true value, it follows logically that the true value is within *t* standard errors from the sample mean. Written in equation form,

$$\text{true value} = \text{sample mean} \pm t \text{ standard errors}$$

Stated another way, we are 95% confident that the true value lies in the interval between the sample mean minus *t* standard errors and the sample mean plus *t* standard errors. Written as an equation, we have

$$\overline{X} - t\frac{S}{\sqrt{n}} \le \mu \le \overline{X} + t\frac{S}{\sqrt{n}}$$

The interval given in the above equation is called the *confidence interval.* Thus, a confidence interval is the range of values that are believed to contain the true parameter value. The value of *t* can be found in a

statistical table or with a computer. To simplify the math, Table 10-5 combines the *t* value and the square root of the sample size into a *k* factor. This simplifies the above equation to

$$\overline{X} - kS \le \mu \le \overline{X} + kS$$

Table 10-5. Factors for determining 95% confidence interval for the mean.

n	*k*	*n*	*k*
2	8.98	16	0.53
3	2.48	17	0.51
4	1.59	18	0.50
5	1.24	19	0.48
6	1.05	20	0.47
7	0.92	21	0.46
8	0.84	22	0.44
9	0.77	23	0.43
10	0.72	24	0.42
11	0.67	25	0.41
12	0.64	26	0.40
13	0.60	27	0.40
14	0.58	28	0.39
15	0.55	29	0.38
		30	0.37

n = sample size

For example, suppose we perform an experiment that produced a set of 20 PaO_2 measurements with a mean value of 83 mmHg and a standard deviation of 5 mmHg. We wish to construct a 95% confidence interval (CI) that contains the true mean value of the population of PaO_2 values from which the sample was taken:

$$CI = \overline{X} \pm \left(\frac{t}{\sqrt{n}} \right) \cdot S$$

$$CI = 83 \pm 0.47 \times 5$$

$$CI = 83 \pm 2.35$$

$$CI = (80.6, 85.4)$$

The confidence interval is interpreted as follows: If the experiment was repeated 100 times and 100 confidence intervals were calculated, then 95 of those intervals would contain the true (but still unknown) population mean. Thus we are only 95% confident that the interval we calculated above contains the true value.

Confidence intervals can be calculated for many different statistics. For example, we could calculate the CI for the difference between two means, such as the PaO_2 before and after oscillating positive expiratory pressure (OPEP) therapy. In this case you would calculate the difference in PaO_2 (PaO_2 after therapy minus PaO_2 before therapy) for each patient. Then, you would calculate the mean difference and the standard deviation of the differences to use in the CI calculation.

Confidence intervals are sometimes used in place of hypothesis tests. If the confidence interval includes the hypothesized mean value under the null hypothesis, then we do not reject the null hypothesis. In the example above, if the CI for the difference in mean values contains zero as a value, then we would conclude that there was no difference in the mean PaO_2 before and after therapy (zero might be the true value of the mean difference).

Error Intervals*

Clinical decisions are often based on single measurements, or at best, a few consecutive measurements that show some consistency. We need to estimate the error of single measurements and judge how much confidence we can place in these estimates. Confidence intervals do not provide this information because they summarize groups of data rather than describe individual measurements. We need analogous intervals, which we will call *error intervals*, that describe the combined effects of systematic and random errors on individual measurements. We can also say something about how much confidence should be placed in the estimate. An error interval is the range of values that is expected to contain a given proportion of all future individual measurements at a given confidence level. For example, we could specify a 95% error interval at the 99% confidence level. This specification means that 95 of the next 100 measurements are assumed to lie within this range. Further, if we repeated the experiment that generated the data 100 times, our assumption would be true for 99 of the error intervals we calculated.

Many studies involve the assessment of new devices or methods compared to existing standards, and many statistical concepts are involved. We will simplify the subject and discuss only three cases. An

* A simplified version, called "agreement intervals" was first described by Bland and Altman (Lancet 1986;1:307).

example of the first case would be when a new batch of control solutions for a blood-gas analyzer is purchased. Laboratory standards require us to verify the manufacturer's specifications for the expected lower and upper limits of individual measured values for the analyte. If an individual measurement falls outside one of these limits, the analyzer may be malfunctioning.

A second case would be the evaluation of the performance of a new blood-gas analyzer by measuring control solutions with known PO_2 values.

An example of the third case might be the comparison of PO_2 values from a new model of blood-gas analyzer with measurements from a currently used analyzer.

Although apparently similar, these three problems are actually different. In the first case, it is necessary to determine a range of values within which any given individual measurement should fall when measuring a known value. This range is known as a *tolerance interval*. In the second case, we wish to determine the range of values for the differences between the known (assumed true) value and the measured value. This range is called an *inaccuracy interval.* In the third case, the true value is not known, so we can only assess the range of values for the difference between one measurement system and another. This range is called an *agreement interval.* In each case, the error interval will have the form:

$$\text{error interval} = \text{bias} \pm \text{imprecision}$$

as described in the section on Inaccuracy, Bias, and Imprecision, in the previous chapter. Error intervals will be wider than confidence intervals because the imprecision of individual measurements is larger than that of group statistics (the standard deviation is larger than the standard error by a factor of $\sqrt{n}$).

Tolerance Interval. Given a set of repeated measurements of the same quantity, we are interested in finding the range of values we might observe with a specified degree of confidence. If the true mean and standard deviation of an infinite number of repeated measurements were known, then a "two sigma" (approximately two standard deviation) tolerance interval would be $\mu \pm 1.96\sigma$. This interval includes exactly 95% of the measurements and we can say this with 100% confidence because we have just made an infinite number of measurements. Of course, μ and σ are really unknowable. Therefore, we must substitute the point estimates $\overline{X}$ and S. Because of the random error involved with estimating μ and σ using $\overline{X}$ and S, the proportion of the population of measured values covered by the tolerance interval is not exact. As a result, the confidence level must be based on the sample size.

The *tolerance interval (TI)* is expressed as:

$$TI = \overline{X} \pm k_1 S$$

where

$\overline{X}$ = the sample mean

S = is the sample standard deviation from an experiment consisting of repeated measurements of a known quantity

k_1 = a factor determined so that we can be 99% confident that the given interval will contain at least 95% of all future measurements (see Table 10-6).

Table 10-6. Factors for determining two sigma error intervals or intervals containing 95% of observed measurements at the 99% confidence level.

n	k_1	*n*	k_1	*n*	k_1	*n*	k_1
6	6.37	30	2.85	54	2.55	120	2.32
7	5.52	31	2.83				
8	4.97	32	2.81	56	2.54	130	2.30
9	4.58	33	2.79				
10	4.29	34	2.77	58	2.52	140	2.28
11	4.07	35	2.76				
12	3.90	36	2.74	60	2.51	150	2.27
13	3.75	37	2.73				
14	3.63	38	2.71	65	2.48	160	2.26
15	3.53	39	2.70				
16	3.44	40	2.68	70	2.46	170	2.25
17	3.36	41	2.67				
18	3.30	42	2.66	75	2.44	180	2.24
19	3.24	43	2.65				
20	3.18	44	2.64	80	2.42	200	2.22
21	3.14	45	2.63				
22	3.09	46	2.62	85	2.40	300	2.17
23	3.05	47	2.61				
24	3.02	48	2.60	90	2.38	400	2.14
25	2.98	49	2.59				
26	2.95	50	2.58	100	2.36	500	2.12
27	2.93						
28	2.90	52	2.56	110	2.34	1000	2.07
29	2.87					infinity	1.96

n = sample size

For example, if our we made 20 repeated measurements of a blood gas control solution and found the average $PaCO_2$ was 30 mmHg with a standard deviation of 2 mm Hg, then the tolerance interval for the solution would be

$$TI = 30 \pm 3.18 \times 2$$
$$TI = 30 \pm 6.36$$
$$TI = (23.64, 36.36)$$

Now we could take these results into the blood gas lab and instruct the technicians to do a diagnostic procedure on the analyzer anytime they observe a value below 23.6 or above 36.4 when that control solution level is analyzed.

Inaccuracy Interval. A tolerance interval gives only the imprecision (ie, random error) of measurements. To assess the total error or inaccuracy of a measurement system, we need estimates of both bias and imprecision from experiments in which *known quantities* are repeatedly measured. Bias is estimated as

the mean difference between measured and known (or assumed true) values. Imprecision is estimated as the standard deviation of the *differences* between measured and true values. Inaccuracy is expressed as the sum of the bias and imprecision estimates. Thus, we can construct an *inaccuracy interval (II)* similar to the tolerance interval:

$$II = \overline{\Delta} \pm k_1 S_\Delta$$

where

$\overline{\Delta}$ = is the mean difference between measured and true values

S_Δ = the standard deviation of the differences. The standard deviation of the differences is the same as the difference between standard deviations. The standard deviation of the true value is zero because it is the same value each time. Therefore, we can substitute the standard deviation of the measured values, S, for S_Δ.

k_1 = a factor determined so that we can be 99% confident that the given interval will contain at least 95% of all future measurements (see Table 10-6).

Agreement Interval. How do we know if some new measurement system will give results comparable to the one that is currently in use? Can the new system serve as a substitute for the old while preserving the quality of care? Assessing agreement between two measurement systems is similar to assessing inaccuracy, except that instead of repeatedly measuring on a level of a known quantity, several levels of the known quantity are selected and split samples are measured by both systems. For example, in assessing the agreement between old and new models of a blood-gas analyzer, several samples of patient blood would be split in half and analyzed with each device. Then an *agreement interval (AI),* is expressed as:

$$AI = \overline{\Delta} \pm k_1 S_\Delta$$

where

$\overline{\Delta}$ = the mean difference between the pairs of split sample values. The difference for a pair is the result from new machine minus result from old machine.

S_Δ = the standard deviation of the differences

k_1 = a factor determined so that we can be 99% confident that the given interval will contain at least 95% of all future measurements (see Table 10-6).

For example, suppose we wish to compare a new "point of care" blood gas analyzer with our standard "bench" model. We conduct an experiment in which we obtain samples of blood from 100 patients who have a wide range of PaO_2 values. Each sample is split in half, with one portion analyzed on the new blood gas machine and the other half analyzed on the old machine. For each pair of results, we calculate the difference in PaO_2. Then we calculate the mean and standard deviation of the 100 differences. Suppose that $\overline{\Delta} = 5$ mm Hg and $S_\Delta = 9$ mm Hg. The agreement interval is then:

$$AI = 5 \pm 2.36 \times 9$$
$$AI = 5 \pm 21.24$$
$$AI = (-16.2, 26.2)$$

Notice that we rounded the answer to one decimal place because that is the limit of the blood gas machine resolution. We interpret the agreement interval as follows, keeping in mind that the differences were calculated by subtracting the value of the old machine from that of the new machine: Any individual measurement with the new machine can be expected to be anywhere from 16.2 mmHg below to 26.2 mmHg above the value you would have gotten if you had used the old machine. The next question is: Will this expected level of agreement change the quality of medical decision? In this case, you might conclude that the new "point of care" analyzer would lead you to make an unnecessary FiO_2 change on a patient: you would not have made the change if the sample were analyzed on the old machine. Assuming you had confidence in the accuracy of the old machine, you would conclude that adopting the new technology would degrade your standard of care.

Incorrect Methods to Evaluate Agreement. In the past, the most frequently used statistical analyses to assess the comparability of measurement devices were least-squares linear regression, the *t* test, the *F* test, and Pearson's product moment correlation coefficient. However, these techniques are inappropriate for agreement studies.

As mentioned above, least-squares regression minimizes the sum of squares of the vertical distances between the observed data and the regression line. The underlying assumption is that only the data plotted on the *Y* axis show variability. This assumption is inappropriate in the comparison of two measurement systems that both have variability. Also, regression is sensitive to nonlinearity of the data and will give misleading results if the range of data is too narrow.

Both the *t* test (for differences between means) and the F test (for differences between standard deviations) are sometimes used as indicators of agreement. But they are only intended to indicate whether the differences between the two methods are significant. If the calculated value for the statistic is larger than some critical value, the performance of the new system is judged not acceptable. If the statistic is smaller, the conclusion is usually that the methods agree and the new system is accepted. Such judgments may be erroneous for several reasons.

The F test is simply the ratio of the variances of the data from two measurement systems. It is a comparison of error levels, indicating the significance of any difference, and not an indicator of the acceptability of errors or their magnitude.

The *t* test is a ratio of systematic and random errors:

$$t = \frac{bias \times \sqrt{n}}{S_\Delta}$$

where

bias = the difference between the true value and the mean value of the sample ($\overline{X} - \mu$)

S_Δ = the standard deviation of the differences between paired measurements

n = the sample size

As a ratio of errors, *t* does not provide information about the total error or magnitude of disagreement. This situation is analogous to the determination of blood pH by the ratio of bicarbonate to carbon dioxide tension. A low pH does not make clear whether the bicarbonate is low or the carbon dioxide is high. Treatment of acidosis requires information about both metabolic and respiratory errors separately. In addition, there are at least four situations that can cause erroneous judgments when using the *t* value.

- The *t* value may be small when systematic error is small and random error is large. Thus, the farther apart the pairs of measurements are, the more likely we are to conclude the methods agree!
- The *t* value may be small (and we conclude that the methods agree) even when both systematic and random errors are large.
- The *t* value may be large (and we conclude that the methods do not agree) even when both systematic and random errors are small.
- The *t* value gets smaller as the sample size gets larger. Thus, even if systematic and random errors are acceptable, we might erroneously conclude that the methods do not agree if the sample size is large and vice versa.

Perhaps the most widely misused indicator of agreement is the Pearson *r*. The fundamental problem with this statistic is that it is a measure of linear association, which is not the same as agreement. For example, suppose we plotted the data from two methods that gave exactly the same results. The data would all lie on the line of identity. The *r* value would be 1.0 (i.e., a perfect correlation) and we would naturally conclude that the methods had perfect agreement. However, if one method were out of calibration and gave exactly twice the value of the other, or say twice the value plus 3 units, the data would still lie on a straight line with $r = 1.0$. Obviously, the two methods do not agree, as the regression line would be displaced from the line of identity and would have a different slope (indicating both constant and proportional systematic error). The *r* value is an indicator of random error and is completely insensitive to systematic error.

Other problems are associated with the use of the *r* statistic. Correlation is sensitive to the range of measurements and will be greater for wide ranges than for small ones. Because you will usually compare methods over the whole range of values expected to be measured, a high correlation is almost guaranteed. Also, the test of significance (that *r* is significantly different from zero) will undoubtedly show that the two methods are related, as they are designed to measure the same quantity. The test of significance is therefore irrelevant to the question of agreement.

All of the above statistics share one additional weakness: they all describe characteristics of a group of data and say nothing about individual measurements. Recall that this is important because many clinical judgments are based on single measurements.

Data Analysis for Device Evaluation Studies

Establishing Standards. Measurement system performance studies are intended to show how close a single measurement value will be to the true value and how much confidence can be place in it. Before any judgment can be made, we must have decided *beforehand* what level of inaccuracy is acceptable. This decision can be both elusive and confusing. Standards for allowable error may be generated in several ways:

- On the basis of the intended application. For example, a simple oxygen analyzer used in an adult ICU may have an allowable error of $\pm$ 2 percent of full scale because a small discrepancy will have little clinical effect. However, measurement of oxygen concentration for the purpose of calculating gas exchange parameters requires much better accuracy. For example, an error of 1% in the measurement of oxygen concentration leads to an error of 24% in the calculation of oxygen consumption and 32% in the calculation of respiratory exchange ratio. For this purpose, the allowable error would be about 0.1%.

- On the basis of agreement with similar, commonly used measurement systems. For example, the accuracy of pulse oximeters should be comparable in-vitro oximeters.
- On the basis of professional consensus. For example, the American Thoracic Society has published accuracy and calibration standards for pulmonary function equipment.
- On the basis of arbitrary statistical methods. For example, if the standard deviation of the measurement method is one-fourth or less of the normal population standard deviation, then analytic imprecision may be judged negligible.

Note that even if the allowable error can be agreed upon, we must know the value, or range of values, of measured quantities that represent cutoff points for medical decisions. For example, an imprecision in transcutaneous PO_2 readings of $\pm$ 15 torr might be reasonable for PO_2s above 100 torr but would not be adequate for lower values in the range that might indicate hypoxemia.

General Experimental Approach. In general, pairs of data (measured versus known values or values from two methods) should be gathered over a wide range of input levels (one data pair per level). These data will provide information about bias and some information about imprecision. Additional data pairs may be gathered from repeated measures at selected levels (critical levels corresponding to cutoff points for making decisions) to provide better estimates of imprecision. The sample size for a given experiment will be limited by many practical factors but should be no less than 20, with larger samples being preferable.

Experimental results should be planned with a consideration of the possible sources of bias and imprecision discussed above. For example, to estimate the inaccuracy expected in the normal daily operation of a measurement system, data should be collected over the entire range of measurements that will be used clinically (to account for errors due to the magnitude of measurements) and over a period of days (to account for environmental and operator factors and even calibration errors). On the other hand, if the inaccuracy of the system alone is desired (to compare it with another system) repeated measures of the same quantity should be made within a short period, with the same operator, and with all other possible confounding factors held as constant as possible.

Data Analysis Procedure

Step 1. The first step in any data analysis should be to create a scatter plot of the raw data to get a subjective impression of their validity. The mean and standard deviation should be calculated and used to identify outliers (see below). In determining *tolerance intervals*, the data should be plotted along with the data mean and standard deviation. For *inaccuracy intervals*, the difference between the true (known or standard) values and the measured values should be plotted on the vertical axis against the true values on the horizontal axis. For *agreement intervals,* the differences between data pairs from the two measurement systems are plotted against the mean value for each pair. The mean values are used as the best estimate of the true values, which are not known. The purpose of these plots is not only to identify any outliers but also to make sure that the differences are not related to the measurement level. If they are, the standard deviation may be overstated at one point of the measurement range and understated at another. In other words, calculation of one value for standard deviation for all the data will not accurately describe the variability of the data over the entire range of measurements. You might be able to solve this problem by using a logarithmic transformation of the data. Alternatively, you could derive worst-case error specifications as a percentage of full scale. The hypothesis that the data points are correlated with the measurement level can be tested formally with Pearson's product-moment correlation coefficient (ie, test the hypothesis that r is significantly different from zero).

Step 2. The next step is to make sure the data comply with the assumption of normality (that they are adequately described by the normal or Gaussian distribution). All of the statistical procedures described hereafter are based on this assumption. The data can be assessed for normality with the Kolmogorov-Smirnov test. It may be sufficient to simply plot the frequency distribution and make a subjective judgment as to whether or not it is "bell-shaped".

Step 3. Once the data are judged to conform to the underlying assumptions, the mean and standard deviation are used to calculate error intervals as described above.

Step 4. Finally, the data should be presented in graphic form and labeled with the numerical values for the error intervals (Figure 10-14).

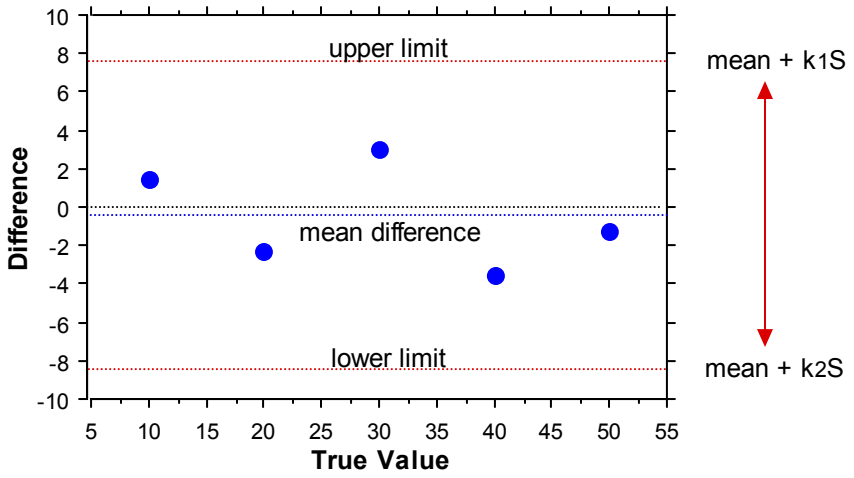


Figure 10-14. Suggested format for plotting data and error intervals for inaccuracy and agreement studies. Each data point represents the difference between the measured value and the true value (inaccuracy study) or the difference between two measurements of the same specimen using two different devices (agreement study). For inaccuracy studies, the magnitude of the difference (vertical axis) is plotted against the true or known value (horizontal axis). For agreement studies, the true value is not known so it is estimated as the mean of the two measurements (horizontal axis). For example, S = sample standard deviation and k_1 is the factor for determining a two-sigma error interval (from Table 10-6). The product of k_1 times S is added to the sample mean to get the upper limit of the error interval. The product is subtracted from the mean to get the lower limit.

If the study is designed to compare the error intervals of several measurements, the data can be plotted as shown in Figure 10-15.

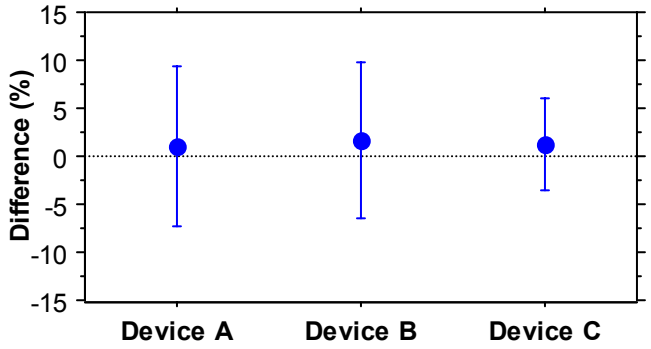


Figure 10-15. Suggested format for plotting error intervals when comparing several different devices (accuracy study). The dots represent the mean difference between measured and true values. The vertical lines through the mean values represent the error intervals calculated as the mean difference $\pm$ k_1S as shown in Figure 10-14.

Treatment of Outliers. Spurious data may be caused by temporary or intermittent malfunctions of the measurement system or by operator errors. Such errors should not be included in the error analysis. The plot of the data should be inspected for outliers, or values that depart from the expected distribution. Outliers at the upper and lower limits of the range will have a strong effect on the estimates of systematic error. Any of the sources of error mentioned previously can cause unusually large deviations from the desired linear relation and falsely increase the estimate of inaccuracy. Outliers are identified as any measurements that are more than k_2 standard deviations from the mean, where k_2 is based on the sample size (Table 10-7). Any occurrence of an outlier should be examined for evidence of a real source of nonlinearity rather than assuming it to be a spurious error. The occurrence of more than three unexplained outliers per 100 observations suggest the presence of a serious problem with the measurement system. Outliers are eliminated from the data one at a time until no more are identified. Each time an outlier is rejected, a new mean and standard deviation are calculated for the reduced sample (Figure 10-16).

Table 10.7. Factors for determining outliers at the 95% confidence level.

n	k_2	n	k_2
3	1.15	13	1.84
4	1.39	14	1.85
5	1.57	15	1.86
6	1.66	16	1.87
7	1.71	18	1.88
8	1.75	20	1.89
9	1.78	25	1.90
10	1.80	30	1.91
11	1.82	40	1.92
12	1.83		

Interpreting Manufacturers' Error Specifications

In evaluating a new device, our main concern is "how much error can be expected in normal use". And knowing that any specification of error is just an estimate, we want to know how much confidence to place in it. You would think this is a straightforward question with a simple answer. But manufacturers can be rather cryptic about their error specifications. For example, the operator's manual for a leading brand of pulse oximeter has a section titled "Specifications." There we find a specification for "accuracy" listed as " percent $SpO_2 \pm 1$ SD". From this specification we gather that the error in measured saturation is given as a "one-sigma" interval (one standard deviation). Next, we find that for saturations in the range of 70 to 100 percent, the accuracy is "$\pm$ 2 digits". Given that the readout is digital and the units for SpO_2 are percent, we surmise that "digits" means percent. So what does this mean? First, we have to assume that the standard deviation (SD) referred to is that of the population of all future measurements made with the instrument, because we are not given a sample size. (Besides, we happen to know that error specifications for pulse oximeters are based on thousands of measurements, so the sample is a good estimate of the population.)

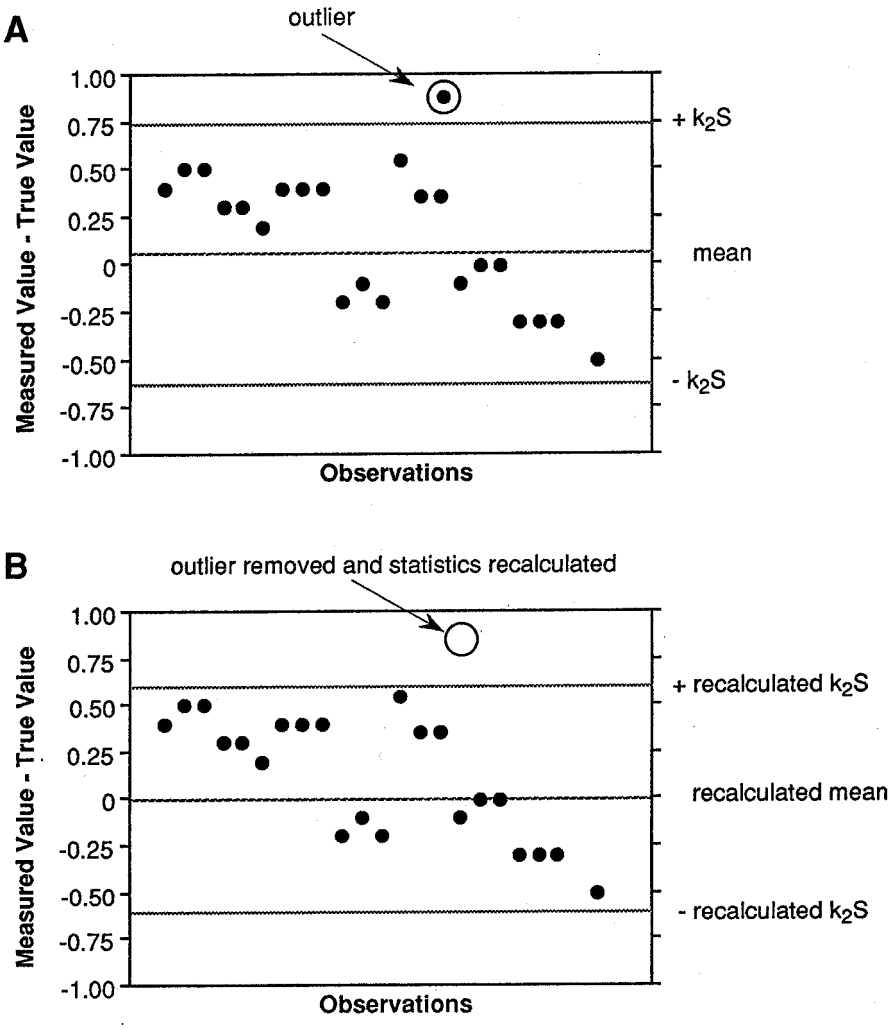


Figure 10-16. Procedure for handling outliers in a set of measurements. An outlier is identified in the original data set as a data point outside the limits calculated as $\pm$ k_2 standard deviations (see Table 10-7). It is removed and the outlier limits are recalculated. This time, no outliers are found.

Second, as clinicians, we like to keep our risk of error below the 5 percent level for any kind of measurement. To achieve that level, we have to double the specified error to get a two-sigma interval (see Figure 10-11). Thus, we can expect 95% of measurements with this device to fall within $\pm$ 4% of the true value. And because we have no further information, we just assume that we can have 100% confidence in this estimate.

Next, consider the operator's manual for a leading manufacturer of serum ionized calcium analyzer. The manufacturer has defined inaccuracy as the mean difference between the measured value on a group of instruments and the estimated true value (what we have called bias). Imprecision is defined as the standard deviation of the measurements. You will recall that the standard deviation of the measurements is the same as the standard deviation of the differences between measured and true values, assuming the true value is constant. We are told that 100 samples were analyzed with a calculated bias of 0.04 mmol/L and imprecision of 0.02 mmol/L. We have enough information to create a two-sigma error

interval at the 99% confidence level (an inaccuracy interval). From Table 10-6 we see that the k_1 value for $n = 100$ is 2.36. The inaccuracy interval is thus

$$0.04 \pm 2.36 \times 0.02 = 0.04 \pm 0.05 = (-0.01, 0.09)$$

Note that the observed inaccuracy of a device may be different from the manufacturer's specifications, depending on how it is used. Instrument specifications do not, for example, include the various types of operator errors.

The user must be aware of the implications of error specifications. If a specification is given as a percentage of the full scale reading, the device will be more accurate for measuring quantities at the upper end of the scale than on the lower end. For example, suppose the scale range is from 0 to 100 and the error is $\pm$ 2 percent of full scale. If a known quantity having a value of 95 is measured, the instrument will probably give a value between 93 and 97, which is 98% to 102% of the true value. However, if a know quantity of 5 is measured, the reading will be between 3 and 7, or 60% to 140% of the true value. This represents an inaccuracy for that particular reading of $\pm$ 40%, which might be unacceptable, depending on the application. On the other hand, if the specification is given as a percent of reading, the instrument will have the same error at the upper and lower ends of the scale.

Inverse Estimation. In *creating* error specifications, the manufacturer's task is to describe the spread of measured values around the true value. But in *using* error specifications, the problem is reversed. For a given measured value, we want to know the range of values in which the true value will lie. If the mean (systematic error), standard deviation (random error), and sample size are given, we simply construct an inaccuracy interval, as described previously.

When the error specification is given as a percentage of full scale, the systematic and random errors have been lumped together into an equivalent constant systematic error. The lower and upper limits for the true value ($true_L$ and $true_U$) are given by:

$$true_U = measured\ value + \frac{error \times full\ scale\ reading}{100}$$

$$true_L = measured\ value - \frac{error \times full\ scale\ reading}{100}$$

For example, an instrument with a scale of 0 to 200 and an error of $\pm$ 4 percent of full scale show a reading of 83. The true value will be in the interval

$$true_U = 83 + \frac{4 \times 200}{100} = 91$$

$$true_L = 83 - \frac{4 \times 200}{100} = 75$$

Notice that the true value is between 8 units below and 8 units above he measured value; the error band is symmetrical.

When the error specification is given as a percentage of reading, the systematic and random errors have been lumped together into an equivalent proportional systematic error. Because the error is proportional to the reading, the limits of the estimated true value are not symmetrical. The lower limit is closer to the measured value than the upper limit. In this case, the limits are given by

$$true_U = \frac{measured\ value \times 100}{100 - error}$$

$$true_L = \frac{measured\ value \times 100}{100 + error}$$

Using the same example as before, but with an equivalent error of 10 percent of reading, we get:

$$true_U = \frac{83 \times 100}{100 - 10} = 92$$

$$true_L, = \frac{83 \times 100}{100 + 10} = 75$$

Now the true value is between 8 units below and 9 units above the measured value.

Hypothesis Testing

In hypothesis testing, we translate a research hypothesis into a formal mathematical statement, obtain a sample statistic, and find the probability of our sample statistic from the appropriate sampling distribution. Based on the probability we obtain, we either accept or reject our hypothesis.

Hypothesis testing is a technique for *quantifying* our guess about a hypothesis. We never know the "real" situation. Does drug *X* cause *Y* or not? We can figure the odds, and quantify our probability of being right or wrong. Then, if we can decide what odds (or risk of being wrong) we can live with, we have a ready-made decision rule. If the probability that drug *X* causes result *Y* is high enough, then we decide that *X* does cause *Y*. We first give an illustrated example of a hypothesis test, and then summarize the steps in the procedure. You may have difficulty understanding the concepts of hypothesis testing until you actually work through a problem.

Example. Let's say you are a staff therapist in a busy University hospital. Much of your day is devoted to delivering aerosol medications to patients on acute care floors. Thinking of yourself as a lone star in a world of "slackers", you wonder if you are actually working harder than your colleagues. For one week, you record the number of aerosol treatments you give each day. You find that you average 32 aerosols per day. Then you ask your supervisor what a fair aerosol treatment workload should be. He looks up some historical data and tells you that based on the total number of aerosol treatments done in a year and the number of therapist in the department, the average should be about 26 per day. Now the question is whether your average is significantly higher than the department average, or is your average due to a chance fluctuation? In other words, do you really work harder on average than the rest of the department or were you just having a busy week?

The key concept to remember is that the value of a sample mean will fluctuate with different random samples, even if all samples come from the same population. This fluctuation is due to random sampling error. In other words, even though the population mean is 26, different samples will have different values for their means. A sample mean of 32 is possible even if the population that the sample comes from has a mean of 26. Of course, we don't know whether your sample represents an actual mean of 26 with a chance fluctuation, or represents a truly different population (your average assignment) with a mean greater than 26.

It is possible by chance to have a sample mean of 32 with a true population mean of 26 because of random error. Therefore, we calculate the probability of a sample mean of 32 to make our decision. If the probability is high, then we accept that the sample mean of 32 represents a population with a mean of 26. We conclude that there is no *actual* difference between 26 and 32, only a *chance* difference. This conclusion represents the *null* (no-difference) case or hypothesis. The symbol for the null hypothesis will be H_0. On the other hand if probability of a sample mean of 32 due to chance given a population

mean of 26, is low, then we will reject the no-difference hypothesis and accept the alternate hypothesis that your sample workload comes from population with a mean that is greater than 26, and there is a difference between 26 and 32. The symbol for the alternate hypothesis is H_A. We must now decide what probability values we would consider to be "high" or "low".

If we set a cutoff value, below which all values are considered low, then we have set a *significance level.* The significance level is symbolized by the Greek letter alpha (α). For example, if alpha is set at 0.05, we are saying any probability of 0.05 or less is not likely. The value of 0.05 is commonly used by convention. In the methods section of most research articles you will find a statement like "The significance of our statistical tests was set at 0.05".

Let us summarize the information we have:

- Outcome variable: Workload in units of aerosols delivered per person per day
- Research hypothesis. My workload is higher than the departmental average.
- Sample data: (33, 49, 22, 27, 30)

Now let's review the procedure for testing the hypothesis:

Step 1. Formulate the statistical hypotheses.

H_0: $\mu = 26$ (null hypothesis)

H_A: $\mu \neq 26$ (alternate hypothesis)

The null hypothesis is that your sample came from a population with a mean of 26. The alternate hypothesis is that the sample comes from a population with a mean of greater than or less than 26, since we cannot foretell in which direction it might be. For example, your average workload may indeed be higher than the department's. On the other hand, your average workload may less, but you were having a really busy week when you collected your sample. This type of hypothesis and is called "two tailed" for reasons that will become clear later.

Step 2. Calculate the descriptive statistics for the sample data.

$$\overline{X} = \frac{33 + 49 + 22 + 27 + 30}{5} = 32.2$$

$$S = \sqrt{\frac{(33-32.2)^2 + (49-32.2)^2 + (22-32.2)^2 + (27-32.2)^2 + (30-32.2)^2}{5-1}} = 5.9$$

Of course, you would not do this calculation by hand as shown above but would use a spreadsheet or a calculator. Note that we are retaining one more significant digit in the answers than in the data to prevent round off errors. The final answer will have the correct number of significant digits (no more than the fewest number in the data, in this case, 2). You have to pay attention to little things like rounding error.

Step 3. Calculate the test statistic.

In hypothesis testing, we assume that there is no difference (assume the null hypothesis is true) and find the probability of our sample mean value. To do that, we need to know what probability distribution to use. The appropriate distribution is our sampling distribution. Let us assume our sampling distribution is adequately described by the *t* distribution (because the sample size is small and we do not know the

department's population standard deviation). Now, what is the probability of a sample mean of 32 *or greater* if the population mean is 26?

We calculate the *t* statistic for our data:

$$t = \frac{\overline{X} - \mu}{S/\sqrt{n}} = \frac{32 - 26}{5.9/\sqrt{5}} = 2.374$$

Step 4. Determine the rejection region under the t distribution curve.

We need the "cut-off" value of *t* associated with the desired significance level (0.05). This cut-off value of *t* allows us to rule off two areas under the *t* distribution curve that each represents half of the desired significance level (again, because we cannot say beforehand whether the population our sample came from has a mean higher or lower than the department mean). These areas represent unlikely values of *t* in the "tails" of the distribution. That is why the hypothesis test it is called a "two-tailed" test (Figure 10-17). The idea is that if the value of *t* from our sample data (Step 3) is larger than the critical value of *t*, then we conclude that our results are unlikely to occur by chance.

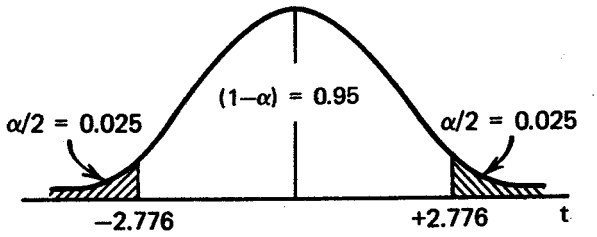


Figure 10-17. Areas under a *t* distribution for a sample size of 5. The *cutoff* value of ±2.776 represents the distance above and below the true population mean that encompasses a probability of 95%. The cutoff value will change for different samples sizes or significance levels. The shaded areas represent the "rejection regions".

We can look up the cutoff value in a statistical table or we can use the Microsoft Excel spreadsheet equation

=TINV(significance level, degrees of freedom)

where the degrees of freedom equals *n*-1. Thus, we type

=TINV(0.05,4)

in a cell, press the enter key and get the value 2.776.

Figure 10-17 is the key to hypothesis testing and confidence intervals. If you can understand all the concepts it illustrates, then you will be prepared to read research articles and even do basic statistical testing.

The graph in Figure 10-17 is somewhat simplified in that there is no vertical axis. The vertical axis is probability, as we discussed in an earlier section. The horizontal axis gives values of *t* from negative infinity to positive infinity. That is why the curve never touches the horizontal axis. The vertical line in the center of the curve marks the mean of the distribution. As with the *z* distribution, the mean of the *t* distribution is zero. The vertical line also corresponds with the mean value of the population we are assuming under the null hypothesis, H_0, which in this example is 26.

The vertical line divides the distribution into two equal areas of 50% each. But we are more interested in the two shaded areas. The shaded areas, added together, make up a probability of 5%. That 5% corresponds to the significance level, α, that we set beforehand. Those areas represent occurrences that we consider rare. The unshaded area represents occurrences that are not rare. Occurrences of what? Well, we mean values of *t*, which represent occurrences of sample statistics (in this case mean values) from experiments. So, a sample mean that was close to the population mean would produce a *t* value close to zero and it would be in the unshaded region. A sample mean that was unusually far away from the mean would produce a *t* value that would be in the shaded region. We would consider the occurrence of such a sample mean (and corresponding *t* statistic) rare. So rare that if we observe one we would sooner conclude that it was from a different curve, a curve with a different mean. That is why the shaded regions are called "rejection" regions, because we would reject the null hypothesis. The unshaded regions are called acceptance or more correctly, "do not reject" regions. (Remember, philosophically speaking, we never really accept that a hypothesis is true; we simply do not reject it until such time as more data are available.)

Step 5. Make a conclusion.

If you have understood everything up to this point, making the conclusion is simple. Your value of *t* from the sample is 2.374. It is between zero and the critical value of 2.776. That puts your *t* value within the "do not reject" region of the curve, so you do not reject the null hypothesis. Translated, that means you cannot conclude that your sample workload came from a population with a higher mean. So, despite the fact that for one week your average workload was higher than the departmental average, you cannot conclude that in general, your productivity is higher than your colleagues. What could have caused the mean to look higher but still not be a significant difference? Look again at the sample data. Notice that there is one day that sticks out at 49 treatments. This relatively large value had two effects. It increased the mean value a little but it increased the standard deviation substantially. For example, if that day had been 32 instead of 49, the mean would be 6 percent lower but the standard deviation would be 59 percent lower. Now look at the equation for *t* (page 165). Notice that as the standard deviation increases, *t* decreases. True, the mean also increases which would tend to increase *t* but we have just seen the effect on the standard deviation is much larger. So data with a lot of variability are less likely to indicate a significant difference. Also notice that *t* is directly proportional to the square root of the sample size. That means our small sample size of only 5 days also tended to make the difference in the mean values insignificant. Consider what would have happened if we set the significance level at 0.10 or 0.01 instead of 0.05.

What would have happened if your sample had produced a *t* value larger than 2.776? That value of *t* would have been in the rejection region. Therefore, you would have concluded that your sample came from some population with a larger mean value than the population the departmental data represented. What you are really saying is that if the sample mean of 32 is unlikely to have come from a population with a mean of 26, then it follows logically that the sample is more likely to have come from a population with a larger mean. We are not saying what that larger mean value is, just that it exists. Figure 10-18 illustrates this reasoning: Under the null hypothesis (population mean = μ_0), observed values far from the mean are rare. If they are rare enough (occur less than 5% of the time) we consider them to be in the rejection region (shaded area un upper curve). If they are in the rejection region we conclude that they actually come from a different population (mean = μ_1), where they have a higher probability of occurrence (shaded area in lower curve).

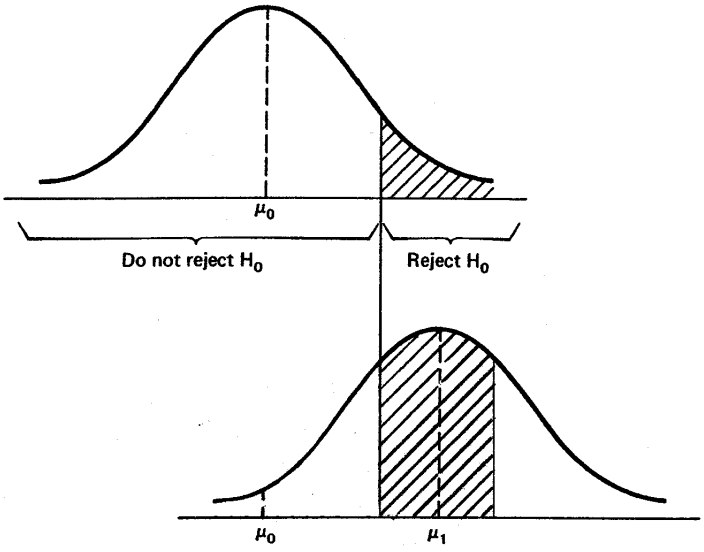


Figure 10-18. Values of the test statistic that lie in the rejection region under the null hypothesis, H_0, (top curve) are assumed to lie in the acceptance region of the distribution described by the alternative hypothesis (bottom curve).

If Figure 10-18 does not make sense to you, consider this simple thought experiment: I place before you a container filled with marbles. I tell you that it contains either 1 white marble and 99 black marbles or 1 black marble and 99 white marbles. Let's say the null hypothesis is that the container is filled with mostly white marbles. You have to test this hypothesis by picking out a sample of five marbles. You do this by removing one marble, noting its color, replacing it, and repeating until you have observed five marbles. If you took an infinite number of samples of five marbles, the expected distribution of the ratio of black to white marbles under the null hypothesis would be represented by the top curve in Figure 10-18. A large percentage of the time you would observe mostly white marbles, represented by the unshaded portion of the curve. A small percentage of the time you would observe mostly black marbles, represented by the shaded portion of the curve.

Now for the experiment: You can't see into the container, just the marbles you take out. You take out one marble, note that it is black, replace it, and shake up the container. You repeat the procedure and again it is a black marble. All five times you draw a marble it is black. What is your guess about the population of marbles? Do you think that you drew the same black marble five times in a row from a container filled mostly with white marbles? That could happen, but it would be very rare. (This corresponds to the shaded region in the top curve of the figure.) Or is it more likely that you were picking from a container with mostly black marbles? (This corresponds to the shaded region in the bottom curve of the figure.) Obvious, your choice would have to be the container of mostly black marbles. We reject the null hypothesis (H_0) and accept the alternative hypothesis. That is all we are doing in Figure 10-18, selecting the distribution that is most probable, knowing that we would make the wrong decision about 5 times in 100 identical experiments (5%) as determined by the significance level of the test.

The step-by-step procedure we have just used can be found in just about any statistical textbook along with tables for looking up *t* values. However, it is not very practical. Assuming you have statistical software or you have set up statistical calculations within a computer spreadsheet, then the procedure is simplified. For example, with your workload data you would do the following:

Step 1. Enter the data into a spreadsheet.

In this case, you simply enter the daily number of aerosol treatments into a single column of the spreadsheet. Statistical software usually has a data entry screen that looks just like an accounting spreadsheet.

Step 2. Select and run the desired statistical analysis.

Your problem requires a single sample *t* test. A statistical program like SigmaStat (www.spssscience.com/sigmastat) even has a "Wizard" that asks you a set of questions about your research problem and then suggests the correct statistical test.

Step 3. Interpret the results.

SigmaStat is very helpful in that it prints out a full report. The report not only summarizes the data but also checks the underlying assumptions of the chosen statistical test. For example, it will check to make sure the sample data are approximately normally distributed. If not, it will suggest an alternative non-parametric test. If the data pass the normality test, you get a calculated value for the statistic, which in our example would be a value for *t*. Then it gives the *p* value. The *p value* is the probability of observing values of *t* equal to or larger than the one for your sample data. In other words, the *p* value is the probability associated with the shaded areas in Figure 10-17. Remember, we set the significance level at 0.05. Therefore, if the *p* value is greater than 0.05, your *t* value does not lie in the rejection region, and you would conclude that there was no significant difference between your sample mean (32) and the population mean under the null hypothesis (the department mean of 26). This is the same thing as comparing the *t* to a critical value and deciding if it is in the rejection region or not. But statistical software and research articles usually give *p* values, not critical values for statistics. For example, using the Microsoft Excel spreadsheet equation:

=TDIST(*t*, degrees of freedom, tails)

where *t* is the value that you calculated from your sample data (2.374), the degrees of freedom = *n*-1 or 4, and tails is 2 for a two tailed test (and 1 for a single tailed test which is seldom seen in the medical literature because it gives less conservative results). Thus, we type:

=TDIST(2.374,4,2)

in a cell, press the enter key and get the *p* value of 0.08. This means that the probability of observing a sample mean of 32 from a population with a mean of 26 would happen 8 times out of 100 experiments, or 8 percent. But our significance level of 0.5 says that only events occurring as rarely as 5 times or less out of 100 are significant. Because 0.08 is greater than 0.05, you conclude that your sample was not sufficiently rare. It probably does not come from a different population.

As a summary and review, important terms used in hypothesis testing are defined.

Research hypothesis: a statement of the proposed relationship between or among variables. For example: T-tube trials reduce weaning time compared to Synchronized Intermittent Mandatory Ventilation (SIMV).

Statistical hypothesis: a precise statement about the parameters of a population. The two forms are the null hypothesis (H_0) and alternate hypothesis (H_A). The null hypothesis states "no difference" or "no association". The alternate hypothesis states that there is a difference or association. For example: H_0:

mean weaning time with T-tube is equal to the mean weaning time with SIMV versus H_A: mean weaning time with T-tube is not equal to the mean weaning time with SIMV.

Test statistic: The statistic, such as a *z* or *t* score, used to test the statistical hypothesis.

Statistical test: a procedure allowing a decision to be made between two mutually exclusive statistical hypotheses, using the sample data. This is also known as a hypothesis test

p value: The probability that the observed results or results more extreme would occur if the null hypothesis were true and the experiment were repeated many times.

Significance level(alpha): The level of probability at which it is agreed that the null hypothesis will be rejected. This is the cutoff level for what is considered improbable.

Type I and II Errors

In hypothesis testing, we have a decision rule for deciding between the two mutually exclusive statistical hypotheses, the null and the alternate. Two decisions are possible: Accept the null hypothesis (the treatment made no difference) or reject it (the treatment was effective). However, we never know for sure the actual situation of whether the treatment is effective.

Two situations are possible in actuality. Either the null hypothesis is true, or the null is not true. This gives four possible combinations between our decision and reality, as illustrated in Figure 10-19.

To simplify the discussion, let us suppose that the actual population mean is either 100 or 103. Both the distribution under the null hypothesis (H_0: mean =100) and the distribution under the alternative hypothesis (H_1: mean = 103) are shown in Figure 10-19. Alpha sets a cutoff point on the distribution under the null hypothesis. Suppose the mean is actually 100 (H_0 is true). Then we have a probability of 1 - α of correctly accepting the null hypothesis. We also have a probability α of rejecting a true null hypothesis. This is called a *Type I error,* the error of rejecting the null hypothesis when it is true. On the other hand, suppose the mean is actually 103, which could be the case as far as we know, and now H_0 is false. Unfortunately, because of random error in sampling, this distribution (right side of Figure 10-19) overlaps the distribution under the null hypothesis. The area of the two distributions that coincides to the *left* of alpha leads to acceptance of the null hypothesis, but the *null hypothesis* is *not true* now. The Greek letter beta (β) symbolizes the probability of a *Type II error*, which consists of accepting a false null hypothesis. Of course if the mean is really 103, the area denoted by 1 - β gives the probability of correctly rejecting the null hypothesis. The term for 1 - β is *power.*

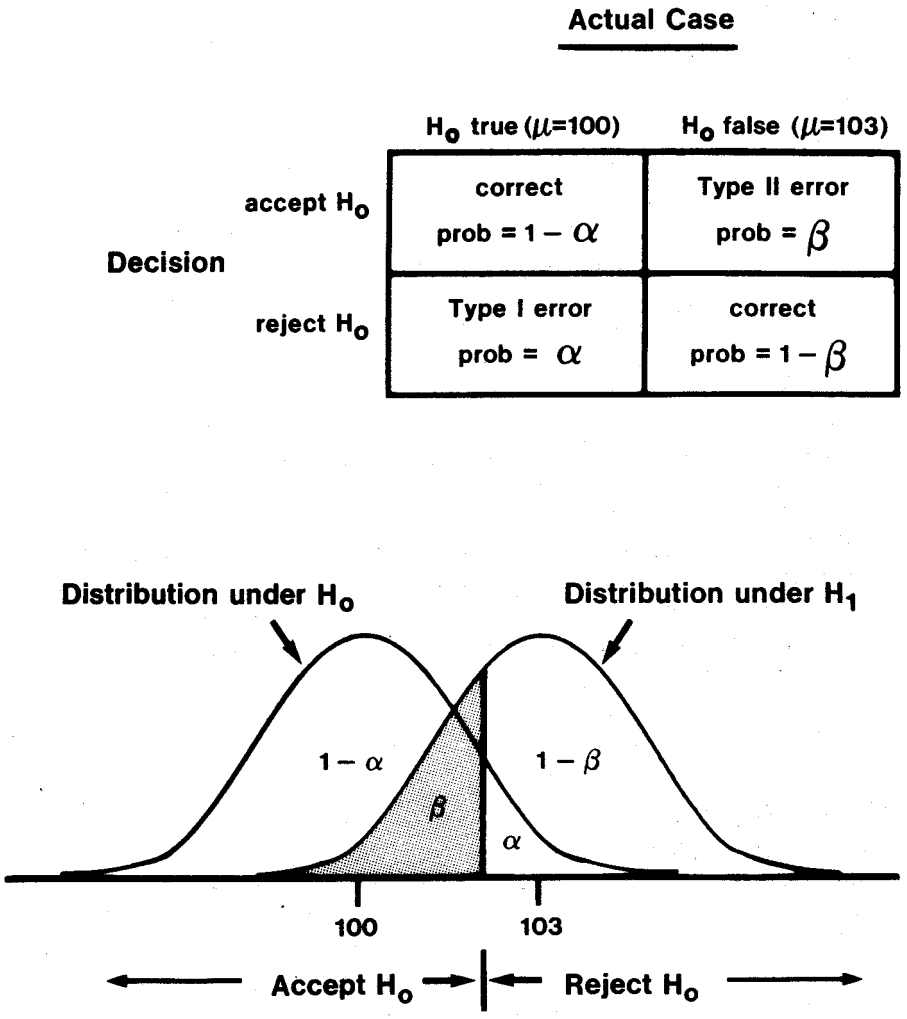


Figure 10-19. Illustration of probabilities for Type I and Type II errors. H_0 = the null hypothesis (for example, mean value = 100), H_1 = the alternate hypothesis (for example, mean value = 103).

Now we can see that lowering alpha to 0.01 or to 0.001 involves a trade-off. If we lower alpha, we reduce the risk of a Type I error, but concomitantly we increase the risk of a Type II error. In Figure 10-19, mentally slide the cutoff point designated by alpha to the right. Alpha is decreased, but the area of beta increases. One solution to keep alpha low *and* to lower beta is to decrease the variability, or spread, of the two distributions. These are sampling distributions, or distributions of the statistic $\overline{X}$. The standard deviation of the $\overline{X}$ values is given by the SEM. A smaller SEM indicates less variation, and we can decrease the size of the SEM, by increasing sample size, *n*, ($SEM = \sigma / \sqrt{n}$). This will decrease the size of beta, and increase power (1 - β).

Which type of error is more serious is relative to the research question. If a new drug is investigated, we would want a definite effect (a large difference) to conclude that the drug is effective. We would desire a small value for alpha, so that we would rather accept a false null hypothesis (Type II error) than reject a true null hypothesis (Type 1 error). In other words, we would rather throw out a drug as ineffective

when it is really effective, than foist a truly ineffective drug on the public, thinking the drug is efficacious. Alpha is made small, although beta increases. On the other hand, if we are testing for accuracy between two monitors, we want only a very small difference before we say the difference is significant. We desire a large alpha, so that a difference is more likely to be called significant. Here we would rather reject a true null hypothesis (Type 1 error) than accept a false, null hypothesis (Type II error). In this case, you would be better off to reject the conclusion of accuracy (even though the monitors are accurate) then say the monitors are accurate when they are not.

Power Analysis and Sample Size

Once the mechanism of hypothesis testing and basic statistical tests are understood, we can discuss the question of adequate sample size, and the related concept of power analysis with statistical tests.

We have previously identified two types of errors that can occur, Type I and Type II (see Figure 10-19). If the null hypothesis represents reality, then α is our probability of rejecting the null hypothesis when it is true (a Type I error), and $1 - \alpha$ is the probability of accepting the null hypothesis when it is true, a correct decision.

But to be complete, we must consider the case where the null hypothesis is false, that is, the alternative hypothesis truly represents reality. Then we saw that there is a probability, β, of accepting the null hypothesis when the alternate hypothesis is really true. The probability of rejecting the null hypothesis, that is, accepting the alternate, is given by $1 - \beta$. The size of β and $1 - \beta$ are determined by the degree of overlap of the sampling distribution under the null and under the alternate hypotheses. This is shown in Figure 10-19.

For a given null and alternate distribution, *decreasing* the risk of a Type I error (lowering alpha) *increases* the risk of a Type II error (beta increases), as we can see in Figure 10-19. By making alpha too small, we can cause beta to become quite large, and thus we run the risk of incorrectly accepting the null. For example, suppose we lowered alpha to 0.001, causing beta to have a value of 0.60. Although we reduced our risk of rejecting a true null hypothesis to 1 in 1,000, we now have a risk of accepting a false null hypothesis that is greater than l out of 2. We could do better tossing a coin instead of doing an experiment! Novice researchers often make the mistake of ignoring power when their study results are negative, meaning that their data apparently showed no difference between outcome variables. If the power of their statistical tests was low, the results must be judged inconclusive.

How do we decrease *both* alpha and beta? Or to put the question another way, how can we have a large probability of accepting a the null hypothesis when it is true ($1 - \alpha$), and also have a large probability of accepting the alternate when *it* is true ($1 - \beta$)? The last probability, $1 - \beta$, is termed *power,* which is the probability of correctly rejecting the null hypothesis.

The most practical means to control power is to manipulate sample size. If we look at Figure 10-19 we can understand the rationale for this. Remember that the distributional curves in Figure 10-19 are distributions of possible sample means. The dispersion in the distributions is given by the standard error of the mean , ($SEM = \sigma / \sqrt{n}$). If we make the SEM smaller, we could decrease the amount of overlap between the two distributions and still have each centered as seen on mean values of 100 and 103. This we can do by increasing sample size.

An example can make the effect of sample size on the SEM, obvious. Let σ equal 9; then for an *n* of 9 and 81 respectively, the SEM is as follows:

$$SEM = 9/\sqrt{9} = 3$$

$$SEM = 9/\sqrt{81} = 1$$

The variance of the population, σ, will not change. Therefore, the SEM will decrease as sample size increases, and the distribution of the sample means will be less dispersed. If the two distributions in Figure 10-19 are each made narrower, then the area indicated by beta must become smaller with a given value for alpha. If beta is smaller, then power, or 1 – β, will be increased.

To summarize, for a given population variance, and alpha level, we can increase power, or the probability of correctly rejecting the null hypothesis, if we use a larger sample. But we may waste time and money by using too large a sample, if a smaller sample achieves a desired power for a given population variance and alpha level.

The power of a statistical test, at any given significance level, is directly proportional to both the sample size and the treatment effect. The treatment effect, or *effect size* is the expected magnitude of the effect caused by the treatment, or independent variable. Effect size is calculated different ways depending on the sample statistic used. However, most often you will think of the effect in terms of a difference between two mean values (such as between two sample means or between a sample mean and a population mean). The difference in means is "standardized" (similar to the reasoning for a z score) by dividing it by the standard deviation. That allows us to use one table or nomogram for any effect size. For hypothesis tests where you will be comparing a single sample mean to the mean of a population, use

$$effect\ size = \frac{\overline{X} - \mu}{S}$$

where

$\overline{X}$ = the sample mean

μ = is the population mean

S = the sample standard deviation

For tests where you want to compare two sample means, use

$$effect\ size = \frac{\overline{X}_1 - \overline{X}_2}{S_P}$$

where $\overline{X}_1$ and $\overline{X}_2$ are the two sample means and S_P is the pooled standard deviation. Once you have the effect size, use the nomogram in Figure 10-20 to estimate either the sample size or the power of the test.

Use the nomogram by connecting any two known values by a straight line and read the unknown value where the line intersects the appropriate scale. For example, if the effect size is 1.0 and the desired power is 0.80, then the required sample size is 30. You can also evaluate the power of a test after the fact. For example, in the experiment we used in the discussion about hypothesis testing, you collected workload data for 5 days and found that the mean workload did not differ from the department's historical average. Because the p value from the hypothesis test was higher than 0.05, we did not reject

the null hypothesis. In doing so, we run the risk of committing a Type II error, accepting the null hypothesis when it is really false. From our discussion about the power of a test, you should suspect that the negative results may be due to the small sample size.

To evaluate the power of the test, we first need to calculate the effect size. We can start with the actual mean values. You will recall that your sample mean from 5 days was 32.2 with a standard deviation of 5.9, while the population mean from departmental records was 26:

$$effect\ size = \frac{\overline{X} - \mu}{S} = \frac{32.2 - 26}{5.9} = 1.05$$

On the nomogram in Figure 10-19, we line up the effect size of 1.05 with the sample size of 8 (the smallest number on the scale and fairly close to 5) and extend the line through the power scale. Where the line intersects the scale we read a power of about 0.33. From Figure 10-19 we see that the probability that we made a correct decision was only 33%. Because power = 1 – β, then the probability of a Type II error is β = 1 – power. Thus, the conclusion that your workload was no different from the departmental average had a probability of 1 –0.33 = 0.67, or a 67% chance of being wrong! That does not mean that you are great and the other workers are slackers, it means that you did not collect enough data to make a strong conclusion.

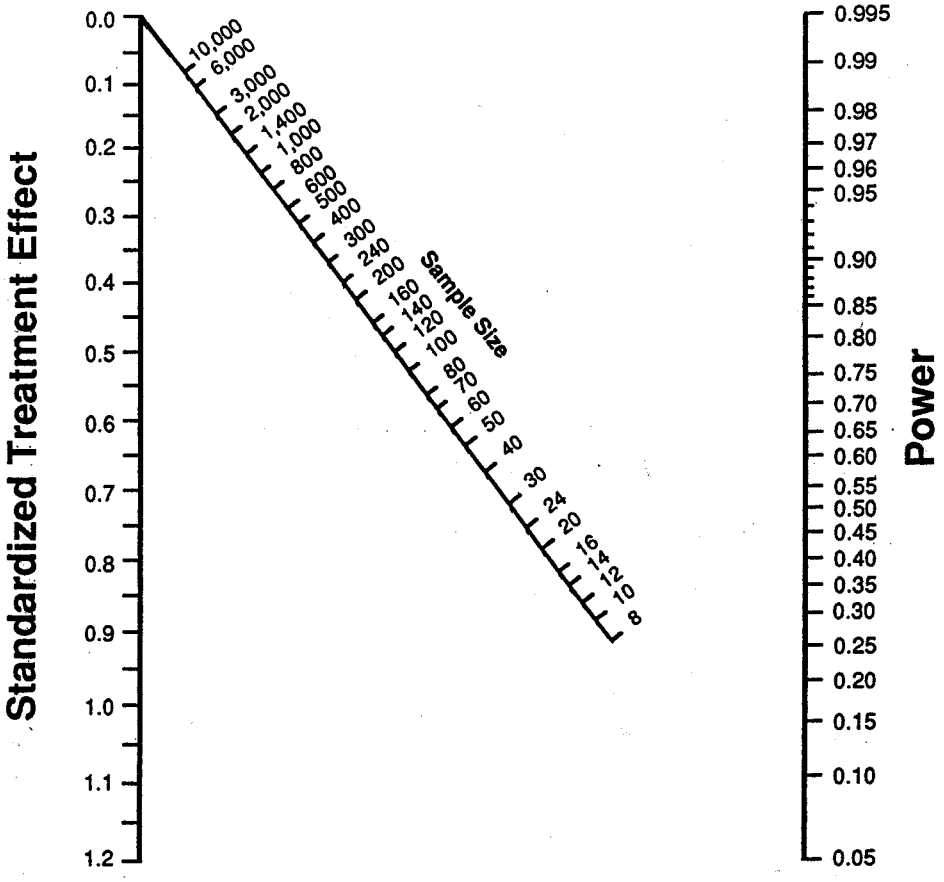


Figure 10-20. Nomogram relating effect size, total study size (sum of two equal sample sizes), and the power of a *t* test for the difference between mean values. Connect any two known values by a straight line and read the unknown value where the line intersects the appropriate scale. For example, if the effect size is 1.0 and the

desired power is 0.80, then the required sample size is 30. If the study requires only one sample, divide the sample size by 4 for a rough approximation.

The selection of power is somewhat arbitrary, depending on which type of error is more serious in a given research situation. Usually, Type I error is more serious than Type II error: We would not want to conclude that there *is* a treatment effect if one does not exist because the subject is unnecessarily exposed the risk of adverse reactions to treatment. Usually, a power of 0.80 (β of 0.20) is commonly used.

So how many days of data would you have to collect to be confident in your results? First decide on how big of an effect you want to show. Let's say that you would be content to show that you worked 20 percent harder than your co-workers. Twenty percent of 26 is 5.2. Now calculate the effect size:

$$effect\ size = \frac{5.2}{S} = \frac{5.2}{5.9} = 0.88$$

Second, decide how much power the hypothesis test should have. Again, this is an arbitrary decision; based on how much you want to avoid making a Type II error (low power means high probability of error). Concluding that your workload is no different from your coworkers will cause little more harm than hurting your ego, so you set the power at 0.75. Now, on the nomogram in Figure 10-20, use a ruler to line up the effect size of 0.88 with the power of 0.75 and you see the required sample size is about 35. Assuming you worked 5 days a week, you would have to collect data for 7 weeks to make an accurate judgment about your workload compared to the average department workload.

There are some important lessons here. First, negative results are not always negative. Your confidence in a negative result depends on the power of the hypothesis test. Second, your efforts in collecting data for five days were not wasted just because you could not make conclusive results. What you did was conduct a *pilot study*, in which you learned something about how much effort is required to collect the data (giving you an idea of the feasibility of the study). More importantly, you were able to estimate the sample variability (the sample standard deviation), and use this to estimate the sample size needed for a larger study that would give more conclusive results.

Rules of Thumb for Estimating Sample Size*

The following rules of thumb let you estimate approximate sample sizes under a variety of conditions without the use of nomograms, tables, or computers. Assume a two tailed hypothesis, significance level = 95% and power = 80%.

Estimates based on mean and standard deviation

These rules are useful when you have some data from a pilot study.

Difference between means (single sample): Use this equation when a single sample mean is compared to a hypothesized population mean,

* Adapded from: van Belle G, Millard SP. STRUTS: Statistical Rules of Thumb. Seattle: University of Washington, copyright 1998. Reproduced with permission.

$$n = \frac{8}{\Delta^2}$$

where Δ is the standardized effect size:

$$\Delta = \frac{\overline{X} - \mu}{S}$$

in which

n = required sample size

$\overline{X}$ = the sample mean

μ = the hypothesized population mean

S = the sample standard deviation

For example, if the standardized effect size is 0.5, then n = $8/0.5^2 = 32$.

Difference between means (two samples): Use this equation when comparing the difference between two sample means:

$$n = \frac{16}{\Delta^2}$$

where Δ is the standardized effect size:

$$\Delta = \frac{\overline{X} - \mu}{S}$$

in which

n = required sample size per group

$\overline{X}$ = the sample mean

μ = the hypothesized population mean

S = the average of the two sample standard deviations (or just take the larger one for a conservative estimate of n)

For example, if the standardized effect size is 0.5, then n = $16/0.5^2 = 64$.

Estimates based on proportionate change and coefficient of variation

These rules are useful when you do not have data from a pilot study but you can say, for example, that you want to detect a 20% change in the mean and the sample data will probably have a 30% variability. You estimate the effect size in terms of a proportionate change of the mean (difference between means divided by population mean) and estimate the variability of the data in terms of the coefficient of variation (population standard deviation divided by the population mean).

Difference between means (single sample): Use this equation when a single sample mean is compared to a hypothesized population mean,

$$n = \frac{4CV^2}{PC^2}[1 + (1 - PC)^2]$$

where

n = required sample size

CV = the estimated coefficient of variation (from the sample data)

PC = the proportional change in the mean you want to detect

S = the average of the two sample standard deviations (or just take the larger one for a conservative estimate of n)

For example, if you would like to detect a 20% change in the mean value and you think the variability of the data should be about 30%, then

$$n = \frac{4(0.30)^2}{(0.20)^2}[1 + (1 - 0.20)^2] \approx 15$$

If you have no idea of what the variability of the data may be, use

$$n \approx \frac{0.5}{PC^2}[1 + (1 - PC)^2]$$

Difference between means (two samples). Use this equation when comparing the difference between two sample means

$$n = \frac{8CV^2}{PC^2}[1 + (1 - PC)^2]$$

where

n = required sample size per group

CV = the estimated coefficient of variation (from the sample data)

PC = the proportional change in the mean you want to detect

S = the average of the two sample standard deviations (or just take the larger one for a conservative estimate of n)

For example, if you would like to detect a 20% change in the mean value and you think the variability of the data should be about 30%, then

$$n = \frac{8(0.30)^2}{(0.20)^2}[1 + (1 - 0.20)^2] \approx 30$$

If you have no idea of what the variability of the data may be, use

$$n \approx \frac{1}{PC^2}[1+(1-PC)^2]$$

Estimates for Confidence Intervals

Frequently we need to calculate a sample size for a fixed confidence interval width. We include the situation where the confidence interval is in the original measurement units (interval width = w) and where the interval is in units of the standard deviation (interval width/standard deviation = w^*).

Confidence interval for the mean

$$n = \frac{16\sigma^2}{w^2} = \frac{16}{(w^*)^2}$$

where σ is the population standard deviation (perhaps estimated from a pilot sample standard deviation).

Confidence interval for the difference between two means

$$n = \frac{32\sigma^2}{w^2} = \frac{32}{(w^*)^2}$$

where σ is the population standard deviation (perhaps estimated from a pilot sample standard deviation).

Sample size for Binomial Test

This rule of thumb is most accurate for sample sizes between 10 and 100.

Difference between proportions (two independent samples)

$$n = \frac{4}{(p_1 - p_2)^2}$$

where p_1 and p_2 are the two proportions and n is the sample size per group. For example, if one sample proportion is 0.5 and the other sample proportion was 0.7, the sample size would have to be 100 for each sample group for the difference to be significant.

Unequal Sample Sizes

Case control. Sometimes a study results in unequal sample sizes. For example, you may not be able to observe more cases of patients with a particular disease but normal controls are plentiful. Suppose n subjects are required per group but only m are available for one group ($m < n$). We need to know how much to increase n to maintain the power of the statistical test. We calculate the factor k such that km is the required larger sample size:

$$k = \frac{n}{2m - n}$$

For example, suppose that the sample size calculations indicate that n = 16 cases and 16 controls are needed in a case-control study. However, only 12 cases are available (m = 12). How many controls will be needed to obtain the same precision? The answer is:

$$k = \frac{16}{2(12) - 16} = 2$$

So we need km = 2 x 12 = 24 controls to obtain the same results as with 16 cases and 16 controls.

Cost control. In some two sample situations the cost per observation is not equal and the challenge is to choose the sample sizes in such a way as to minimize cost and maximize precision. Suppose the cost per observation in the first sample is c_1 and in the second is c_2. How should the two sample sizes n_1 and n_2 be chosen?

$$\frac{n_2}{n_1} = \sqrt{\frac{c_1}{c_2}}$$

This equation is known as the square root rule; pick sample sizes inversely proportional to the square root of the cost of the observations. If costs are not too different then equal sample sizes are suggested (because the square root of the ratio will be close to 1.0). For example, suppose the cost per observation for the first sample is 160 and the cost per observation for the second sample is 40. Then the rule of thumb states that you should take twice as many observations in the second group as compared to the first.

To calculate specific sample sizes, first calculate the required sample size on an equal sample size base using one of the previous equations. Now you know n and k ($k = n_2/n_1$) in the equation from the previous section (Case control). Rearranging that equation and solving for m yields:

$$m = \frac{n(k+1)}{2k}$$

Suppose that on an equal sample basis, n = 16 observations are needed. On the basis of cost, we calculate that we need twice as many of one sample as the other ($n_2/n_1=2$). Then the smaller sample, m, is:

$$m = \frac{n(k+1)}{2k} = \frac{16(2+1)}{2 \times 2} = 12$$

So the sample sizes to minimize cost and maintain precision are 12 and 2 x 12 = 24.

Rule of Threes

The rule of threes can be used to address the following types of question, "I am told by my physician that I need a serious operation and have been informed that there has not been a fatal outcome in the 20 operations carried out by the physician. Does this information give me an estimate of the potential post operative mortality?" The answer is "yes!"

Given no observed events in n trials, the maximum expected rate of occurrence of the event (at the 95 % confidence level) is:

$$rate = \frac{3}{n}$$

In our example, the "observed event" is a fatal outcome. No fatal outcomes have been observed in the last 20 operations. Given no observed events in 20 trials, the rate of occurrence could be as high as 3/20 = 0.15 or 15%. In other words, if the physician had performed 100 operations, we could expect to observe as many as 15 fatalities.

If we know the rate, we can solve for *n*. For example, the rate of extubation failure is 0.15, how many patients will I extubate before I am 95% certain to see at least one fail? The answer is $n = 3/\text{rate} = 20$.

Clinical Importance Versus Statistical Significance

The statistical tests we have discussed have the general form:

$$test\ statistic = \frac{difference}{standard\ error}$$

The size of the test statistic for a given difference is determined by the standard error, which in turn is determined by the sample size. Therefore, statistical significance can be obtained simply by increasing the sample size, so that even a very small effect size is significant. However, we must interpret the results of the hypothesis test with a little common sense. If the difference between two mean values (treatment group vs. control group) is significant but so small that it does not have any practical effect, then we must conclude that the results are not clinically important. For example, when comparing weaning modes, a 20-minute average difference in duration of ventilation may be statistically significant but is it clinically important?

There is admittedly no easy way to determine clinical importance. The sample size should not, however, be artificially increased beyond that needed for an acceptable power level.

Matched Versus Unmatched Data

When picking a statistical test to compare two groups, you must know the relationship (if any) between data points. Data are said to be *unmatched* (or unpaired or independent) if values in one group are unrelated in any way to the data values in the other group. In other words, the values obtained in one group do not affect the values obtained in the other group. The two sample groups are just selected randomly from (supposedly) the same population. In this case, the differences between the two groups are partly due to the effects of the different experimental treatments given and partly due to the natural variability among study subjects.

On the other hand, *matched* (or paired or dependent) data are selected so that they will be as nearly identical as possible. Pairing decreases or eliminates the differences between the two groups due to variability among study subjects. Pairing usually results in the need for smaller sample sizes for any given statistical power. This decreased variability must be accounted for in the statistical procedures used to test hypotheses. Groups of data may be paired in three ways: natural pairing, artificial pairing, and self-pairing.

Natural pairing: Naturally paired groups are obvious pairs such as twins (in humans) or litter mates (in animals).

Artificial pairing: If you cannot get natural pairs, you can get close by selecting pairs of subjects who are matched on as many confounding variables as possible. For example, you could match subjects on age, weight, gender, race, severity of disease, etc. Keep in mind that the more variables you try to match on, the harder it will be to find subjects that match.

Self-pairing: Perhaps the best solution, when possible, is to have each experimental subject act as their own control. For example, you would select a subject, randomly select a treatment, give the treatment, wait until the effects wore off, then give the other treatment. In this way, a single subject generates a pair of data values. This procedure results in the smallest sample size you can get for the desired level of statistical power.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- Qualitative variable
- Quantitative variable
- Discrete variable
- Confidence interval
- Error interval
- Null hypothesis
- Alternate hypothesis
- p value
- Alpha (level of significance)
- Type I error
- Type II error
- Power (statistical)

True or False

1. The number 170.0 has only 3 significant figures.
2. The answer to the calculation: 73.5 + 0.418 should be expressed as 73.9 rather than 73.918.
3. A pie chart is most useful for illustrating the difference between several mean values.
4. A percentiles plot helps you decide how often a certain range of values occurs.
5. The symbol Σ indicates that a set of numbers is to be added together (summed).
6. The bar above the X in the symbol $\overline{X}$ represents the mean (average) value.
7. A correlation coefficient of -0.9 indicates that as one value increases, the correlated value has a strong tendency also increase.
8. The equation for a straight line has the form $Y = a + bX$. The value of a tells you the value of Y when $X = 0$ and the value of b tells you how much Y changes for a unit change in X.

9. The value of r^2 is always larger than the value of r.

10. The power of a statistical test increases as sample size or treatment effect increase but decreases as the sample variability increases.

Multiple Choice

1. Data consisting of categories, such as gender and race, are measured on which level:
    a. Nominal
    b. Ordinal
    c. Continuous (interval)
    d. Continuous (ratio)
2. Data such as height and weight are measured on what level?
    a. Nominal
    b. Ordinal
    c. Continuous (interval)
    d. Continuous (ratio)
3. The Celsius temperature scale is an example of what level of measurement?
    a. Nominal
    b. Ordinal
    c. Continuous (interval)
    d. Continuous (ratio)
4. A pain scale such as None = 0, 1 = moderate, 2 = severe, is an example of what level of measurement:
    a. Nominal
    b. Ordinal
    c. Continuous (interval)
    d. Continuous (ratio)
5. A measure of central tendency (ie, average) appropriate for data measured on the continuous scale is the:
    a. mean
    b. median
    c. mode
6. The ____ is the value below which 50% of the observations occur.
    a. mean
    b. median

c. mode

7. The ____ is most appropriate for data on the nominal level of measurement.
    a. mean
    b. median
    c. mode
8. The statistic that gives you an idea of the average distance from the mean value is the:
    a. range
    b. standard deviation
    c. coefficient of variation
    d. z score
9. The statistic that is calculated as the largest value in a data set minus the smallest value is the:
    a. range
    b. standard deviation
    c. coefficient of variation
    d. z score
10. If you wanted to compare the variability of two different measurements, you would use the:
    a. range
    b. standard deviation
    c. coefficient of variation
    d. z score

You record the number of aerosol treatments given to patients on a particular floor. The resulting data were:

2, 3, 6, 2, 4, 4, 4, 1

Use this set of data for questions 11-15.

11. What is the mean for this set of data?
    a. 3.3
    b. 3.5
    c. 4.0
    d. 5.0
    e. 1.6
12. What is the median?
    a. 3.3

b. 3.5
c. 4.0
d. 5.0
e. 1.6

13. What is the mode?
    a. 3.3
    b. 3.5
    c. 4.0
    d. 5.0
    e. 1.6
14. What is the range?
    a. 3.3
    b. 3.5
    c. 4.0
    d. 5.0
    e. 1.6
15. What is the standard deviation?
    a. 3.3
    b. 3.5
    c. 4.0
    d. 5.0
    e. 1.6
16. The mean ± two standard deviations encompass what percentage of observations?
    a. 68%
    b. 95%
    c. 99.7%
    d. 100%
17. The difference between a confidence interval and an error interval is:
    a. A confidence interval says something about a group of measurements while an error interval says something about individual measurements.
    b. An error interval for a measurement is always smaller than its confidence interval.
    c. Confidence intervals are more useful for interpreting bedside measurements like blood gases.

d. Error intervals are used to test the hypothesis that two mean values are equal.

18. Which error interval is used when the data are composed of the differences between measured and known values?

    a. tolerance interval

    b. inaccuracy interval

    c. agreement interval

19. Which error interval is used to describe the range of values we might expect to find with repeated measurements of a blood gas control solution?

    a. tolerance interval

    b. inaccuracy interval

    c. agreement interval

20. Which error interval is used to compare measurements from two devices when the true value is unknown?

    a. tolerance interval

    b. inaccuracy interval

    c. agreement interval

21. A Type I error occurs when you:

    a. Accept the null hypothesis when it is true.

    b. Accept the null hypothesis when it is false.

    c. Reject the null hypothesis when it is true.

    d. Reject the null hypothesis when it is false.

22. A Type II error occurs when you:

    a. Accept the null hypothesis when it is true.

    b. Accept the null hypothesis when it is false.

    c. Reject the null hypothesis when it is true.

    d. Reject the null hypothesis when it is false.

23. You perform an experiment comparing the effects of two different modes of ventilation on the mean airway pressure. The mean of one group of patients was 12.1 cm H_2O while that of the other was 12.4 cm H_2O. The p value was 0.04. You would most appropriately conclude:

    a. The mean difference in pressure was 0.1 cm H_2O.

    b. There is a statistically significant difference between the groups.

    c. There is no clinically important difference between the groups.

    d. All of the above.

24. Suppose in the example above the p value was 0.40 and the power to detect a difference of 2 cm H_2O was 0.50. Now what would you conclude?
    a. There is a 50% chance of making a Type I error.
    b. The p value indicates that there is no statistically significant difference but the probability of being right in this conclusion is no better than tossing a coin.
    c. The difference is even more significant than before (question 18).
    d. There is a clinically important difference.
25. In the example above (question 19) what could you do to improve the power of the test?
    a. Increase the difference you want to detect to 4 cm H_2O.
    b. Increase the sample size.
    c. Both of the above.
    d. Neither of the above.
26. Matched groups, such as twins, are an example of:
    a. natural pairing
    b. artificial pairing
    c. self-pairing
27. When each experimental subject acts as his own control, the matching is called:
    a. natural pairing
    b. artificial pairing
    c. self-pairing
28. When pairs of experimental subjects are matched on as many confounding variables as possible, it is called:
    a. natural pairing
    b. artificial pairing
    c. self-pairing
29. List all contraindications, adverse effects, unexpected results, and confounding variables.

Chapter 11. Statistics for Nominal Measures

Data on the nominal level of measurement consist of named categories without any particular order to them. Numbers are used here to name, or distinguish the categories, and are purely arbitrary. Nominal measures are usually summarized in terms of percentages, proportions, ratios, and rates. Proportions and percentages are also applicable to ordinal data. Usually, the first step in describing the data is to create a contingency table

DESCRIBING THE DATA

Contingency Table: A contingency table is used to display counts or frequencies of two or more nominal variables. For example, a simple 3 x 2 (rows by columns) contingency table of treatment outcomes might look like this;

Outcome	Treated	Not Treated	Total
Improved	10	1	11
Worsened	2	15	17
No Change	1	2	3
Total	13	18	

Proportion: A proportion is the number of objects of a particular type (such as with a disease) divided by the total number of objects in the group:

$$proportion = \frac{number\ of\ objects\ of\ particular\ type}{total\ number\ of\ objects\ in\ group}$$

For example, if 10 people improved out of a group of 13 people treated, the proportion improved would be 10/13 = 0.77. Thus, a proportion is defined as a part divided by a whole. It is a special case of the mean, where objects are given values of 0 (e.g., for death) and 1 (e.g., for lived). Then the numerator of the equation for the mean is the sum of the 1s and 0s while the denominator is the count of all 1s and 0s.

Percentage: A percentage is a proportion multiplied by 100%:

$$percentage = \frac{number\ of\ objects\ of\ particular\ type}{total\ number\ of\ objects\ in\ group} \times 100\%$$

For example, if 15 people got worse out of a group of 18 people who were not treated, the percentage that got worse would be (15/18) x 100% = 83%.

Ratio: A ratio is the number of objects in a group with a particular characteristic of interest (e.g., died) divided by the number of objects in the same group without the characteristic (e.g., did not die):

$$ratio = \frac{number\ of\ objects\ with\ characteristic}{number\ of\ objects\ without\ characteristic}$$

For example, if a survey of 31 people with a particular disease showed that 13 were treated and 18 not treated, the ratio would be 13/(31-13) = 13/18 = 0.72.

Odds: A ratio of the probabilities of the two possible states of a binary event is called the *odds.*

$$odds = \frac{probability\ of\ event\ occurring}{probability\ of\ event\ not\ occurring}$$

For example, the odds of randomly selecting an ace out of a deck of Poker cards is (number of aces)/(number of remaining cards) = 4/48 = 1/12.

Rate: Strictly speaking, a rate is the number of objects (or quantity of something) occurring per unit of time. However, in statistics, a rate is often defined as:

$$rate = \frac{number\ of\ events\ in\ a\ specified\ period}{total\ number\ of\ events\ in\ specified\ period} \times base$$

where the base is a number used to convert the rate to a conveniently large whole number. For example, if the proportion of deaths in a study was 10/50 and this result was to be related to the whole population, then an appropriate rate might be (10/50) x 1,000 = 200 deaths per 1,000 patients or equivalently, 2,000 deaths per 10,000 patients.

CHARACTERISTICS OF A DIAGNOSTIC TEST

Perhaps the most common source of nominal data is diagnostic tests. The underlying measurements are often at the continuous level of measurement (such as blood gases) but they are used to classify patients as having or not having a condition of interest (like respiratory failure). Given that the condition of interest is either present or absent and the diagnostic test is either positive (condition is present) or negative (condition is absent), four distinct outcomes are possible (Figure 11-1).

Test Result	**Confirmed Condition**	
	Present	Absent
Positive	a True Positive	b False Positive
Negative	c False Negative	d True Negative

Figure 11-1. Characteristics of a predictive test. True positive means test is positive and condition is present; false positive means test is positive but condition is absent; false negative means test is negative but condition is present; true negative means test is negative and condition is absent.

For example, consider a simple bedside screening procedure to predict a patient's ability to be weaned from mechanical ventilation. To date, the procedure most successful at prediction is based on the

rapid/shallow breathing index (RSBI). This index is the ratio of a patient's breathing frequency divided by the tidal volume while breathing spontaneously during a short discontinuation of mechanical ventilation. A ratio of 0.64 has been shown to accurately predict successful continuation of spontaneous breathing without the ventilator for adults.

Suppose you wanted to repeat this study in children to determine if the same cutoff value applies. In addition, you also want to evaluate another index, the CROP index, which is composed of measures for lung compliance, inspiratory pressure, oxygenation and respiratory rate. Ultimately, you want to know which index is more useful and what the cutoff value is.

The experimental data are illustrated in Table 11-1. Notice how data on the continuous level of measurement (RSBI and CROP) have been simplified to the nominal level of measurement (success or failure) so that we can use the statistical techniques designed for diagnostic tests.

Table 11-1. Data for experiment to compare weaning indicators. RSBI = rapid shallow breathing index, CROP = index using compliance, respiratory rate, oxygenation index and inspiratory pressure.

Outcome	RSBI	CROP	Outcome	RSBI	CROP
Success	12.0	0.36	Failure	17.2	0.05
Success	10.1	0.50	Failure	16.7	0.06
Success	9.1	0.52	Failure	13.5	0.13
Success	7.2	0.43	Failure	12.4	0.07
Success	6.3	0.19	Failure	11.6	0.03
Success	5.8	0.48	Failure	9.1	0.04
Success	5.2	0.15	Failure	15.1	0.09
Success	2.4	0.08	Failure	15.0	0.06
Success	3.2	0.45	Failure	11.5	0.08

The next step is to look at the data graphically. We are trying to decide if there is a natural grouping of scores that would discriminate between weaning success and failure.

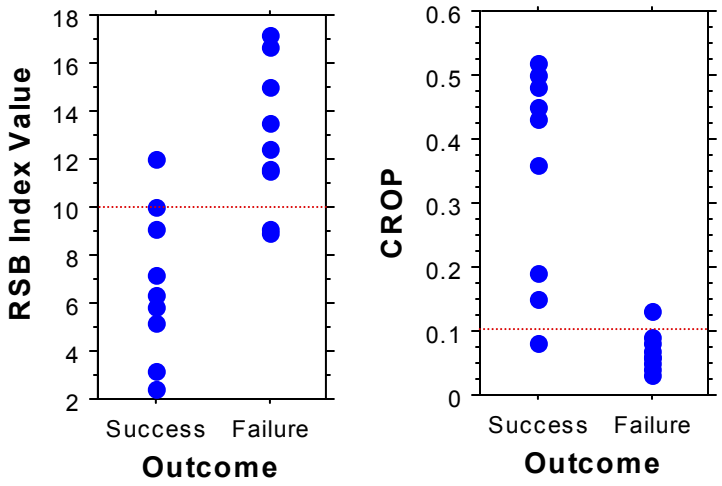


Figure 11-2. Graphical representation of data in Table 11-1. The dotted lines indicate cutoff values that maximize both sensitivity and specificity. For RSBI, values below 10 predict weaning success. For CROP, values above 0.10 predict weaning success.

What we would like is a specific cutoff value for the RSBI such that anytime we observe a patient with a value higher we will conclude that weaning would fail and we continue mechanical ventilation. We would like the same type of cutoff value for the CROP index, but here a high value indicates success. Unfortunately, the plots show that there are no values of RSBI or CROP that perfectly discriminate between success and failure. Therefore, we try to select cutoff values that minimize the false positive and false negative decisions. A false positive decision means that we predict success but the patient fails. A false negative means we predict failure but the patient could really have succeeded if given the chance. We can select the cutoff values mathematically, but we will not go into that here. Assume the dotted lines in Figure 11-2 are the cutoff values for RSBI and CROP, based either on these data or obtained from some other set of data. Now we would like to know how accurate our predictions will be if we use these values to evaluate patients in the future. We evaluate this by converting the data in Table 11-1 into a table that looks like the one in Figure 11-1 using the cutoff values in Figure 11-2. This is shown in Table 11-2.

Table 11-2. Table for calculating the characteristics of the RSB index using the data from Table 11-1 and the cutoff values from Figure 11-1.

		Weaning Outcome	
		Success	**Failure**
Index Prediction	**Success** (positive)	7	1
	Failure (negative)	2	8

True and False Positive Rate

is the probability that the test will be positive when the condition of interest (e.g., ability to wean successfully in our example, or for a lab test, disease) is present. The true positive rate is the same as *sensitivity*. From Table 11-2:

$$true\ positive\ rate = a/(a+c) = 7/9 = 0.78 = 78\%$$

Therefore, out of 100 patients who are weaned successfully, 78 will have an RSBI score predicting success.

The false positive rate is the number of false positive results expressed as a percentage of all positive results. From Table 11-2:

$$false\ positive\ rate = b/(a+b) = 1/8 = 0.13 = 13\%$$

True and False Negative Rate

is the probability that the test will be negative when the condition of interest is absent. The true negative rate is the same as *specificity*. From Table 11-2:

$$true\ negative\ rate = d/(b+d) = 8/9 = 0.89 = 89\%$$

This means that out of 100 patients who failed to wean, 89 will have an RSBI score predicting failure

The false negative rate is the number of false negative results expressed as a percentage of all negative results. From Table 11-2:

$$false\ negative\ rate = c/(c+d) = 2/10 = 0.20 = 20\%$$

Sensitivity and Specificity

Sensitivity is the ability of a test to correctly identify patients with the condition of interest (e.g., ability to breathe spontaneously in our example, or for a lab test, disease). Sensitivity answers the question "If the patient has the condition, how likely is she to have a positive test?" To remember this concept, think *sensitive to disease.* A highly sensitive test is a good *screening* test because it identifies most of the people who have the condition and only a few who do not. From Table 11-2:

$$sensitivity = a/(a+c) = 7/9 = 0.78 = 78\%$$

This means that out of 100 patients who are weaned successfully, 78 will have an RSBI score predicting success.

Specificity is the ability of a test to correctly identify patients who do not have the condition of interest. Specificity answers the question "If the patient does not have the condition, how likely is she to have a negative test?" To remember this concept, think *specific to health.* A highly specific test is a good *diagnostic* test because it identifies most of the people who do not have the condition and only a few who do. From Table 11-2:

$$specificity = d/(b+d) = 8/9 = 0.89 = 89\%$$

This means that out of 100 patients who failed to wean, 89 will have an RSB score predicting failure.

Sensitivity and specificity are not affected by the *prevalence* of the condition of interest. Prevalence is the proportion of the population affected by the condition.

The selection of a cutoff value often involves a tradeoff between sensitivity and specificity as seen in Figure 11-2. For example, suppose we decide that leaving a patient on the ventilator a little longer than necessary was better than having them fail weaning and be reintubated. We might adjust the cutoff value so that false positives are eliminated. For RSBI, we might select a cutoff value of 8 instead of 10. Now the sensitivity is lower (67%) but the specificity is higher (100%). We have just made the RSBI a less useful tool for screening patients who should undergo a weaning trial but a better instrument for predicting failure.

Positive and Negative Predictive Value

The positive predictive value of a test (or predictive value of a positive test) is the probability that the condition of interest is present when the test is positive. From Table 11-2:

$$positive\ predictive\ value = a/(a+b) = 7/8 = 0.88 = 88\%$$

Therefore, out of 100 patients who have an RSBI score predicting success, 88 will likely be weaned.

The negative predictive value of a test (or predictive value of a negative test) is the probability that the condition of interest is absent when the test is negative. From Table 11-2:

$$negative\ predictive\ value = d/(c+d) = 8/10 = 0.80 = 80\%$$

Out of 100 patients with an RSBI score predicting failure, 80 will likely fail the weaning attempt.

Unlike sensitivity and specificity, predictive values are affected by the prevalence of the condition of interest. For example, if the number of people who successfully weaned in Table 11-2 doubled, the positive predictive value would increase:

$$positive\ predictive\ value = a/(a+b) = 14/(14+1) = 0.93 = 93\%$$

and the negative predictive value would decrease:

$$negative\ predictive\ value = d/(c+d) = 8/(8+4) = 0.67 = 67\%$$

Diagnostic Accuracy

The diagnostic accuracy of a test is the proportion of correct results out of all results:

$$diagnostic\ accuracy = (true\ positives + true\ negatives)/all\ results$$

From Table 11-2:

$$diagnostic\ accuracy = (a+d)/(a+b+c+d) = 15/18 = 0.83 = 83\%$$

Likelihood Ratio

The likelihood ratio for a positive test combines sensitivity and specificity into a single number expressing the odds (probability that an event occurs divided by the probability that it does not occur) that the test result occurs in patients with the condition versus those without the condition. A major advantage of the likelihood ratio is that you only have to remember one number, the ratio, instead of two numbers, sensitivity and specificity. From Table 11-2, the likelihood ratio for a positive test is:

$$likelihood\ ratio = sensitivity/false\ positive\ rate = 0.78/0.13 = 6.0$$

We conclude that a positive test result is 6 times more likely to occur in patients who successfully weaned than in patients who failed.

Receiver Operating Characteristic (ROC) Curve

As discussed earlier, there is usually a trade-off between the sensitivity and specificity of a diagnostic or screening test, depending on what value we select for the cutoff. Therefore, a graph that illustrates this relationship would be helpful. Furthermore, we may want to compare two diagnostic tests to see which would be most useful. The receiver operating characteristic (ROC) curve is a device that fills these needs. ROC curves were developed in the communications field as a way to display signal-to-noise ratios (hence the term receiver as in radio receiver). For a diagnostic test, the true positives could be considered the "signal" and the false positives the "noise". The ROC is a plot of the true positive rate against the false positive rate for a test over a range of possible cutoff values (Figure 11-3). The line of identity (diagonal line) corresponds to a test that is either positive or negative by chance alone. That is, the probability of true positive = probability of false positive. In this case, a coin toss would be easier and just as accurate as calculating the index. Plotted values that lie above the line of identity indicate that the diagnostic test is whose predictive ability is better than tossing a coin while values below the line indicate a test that is worse.

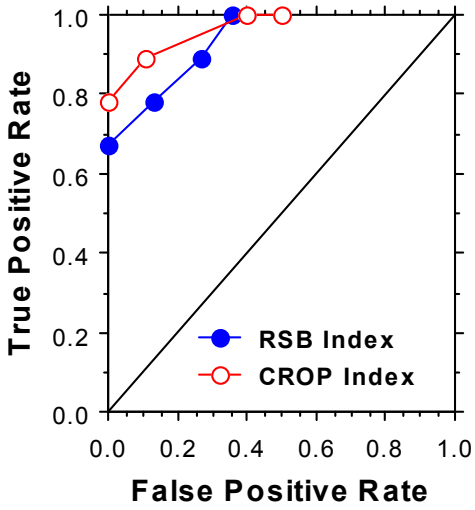


Figure 11-3. Receiver operating characteristic curves for two weaning tests, the RSBI and the CROP index. The CROP is a better test because there is more area under its curve compared to the RSBI curve.

The closer the ROC curve is to the upper left hand corner of the graph (true positive = 100%, false positive = 0%), the more accurate the diagnostic test is. As the cutoff value for the test becomes more stringent (more evidence is required for a positive test), the point on the curve corresponding to the sensitivity and specificity moves down and to the left (lower sensitivity, higher specificity). Note that the false positive rate equals one minus the specificity. If less evidence is required for a positive test, the point on the curve corresponding to sensitivity and specificity moves up and to the right (higher sensitivity, lower specificity). The cutoff value that results in the curve coming closest to the upper left hand corner maximizes both sensitivity and specificity.

Plotting the ROC curves for two diagnostic tests together gives us an easy graphical method to decide which is better. The curve with the most area under it will lie closest to the upper left hand corner over the range of cutoff values and indicates the more accurate test. The area under a the curve for a perfect test is 1.0 while the area for a test that is no better than tossing a coin is 0.5. In Figure 11-3, we see that the CROP index is a more accurate predictor of weaning success than the RSBI for our simulated experiment.

CORRELATION

As discussed in the chapter on basic statistics, correlation means that the values of one variable go up or down with the values of another variable (for example, age is correlated with height). But nominal variables are not "higher" or "lower" when compared, they are just different. Therefore, the idea of association relates to frequencies in categories.

A common issue with diagnostic tests that generate nominal data is to establish the *reliability* of the test. In other words, if the test is repeated, we want to evaluate the agreement between results; high agreement means high reliability. If one person measures the same variable twice and the measurements are compared, an index of within-observer variability is the called an *intrarater reliability* index. When two or more people measure the same variable and their measurements are compared, an index of between-observer variability is called an *interrater reliability* index. Two examples of reliability indexes are the Kappa and Phi coefficients.

Kappa

Frequently in medicine, clinicians must interpret test results as indicating or not indicating disease or abnormality. The outcome is either yes or no, a nominal measure. Suppose we want to evaluate the interrater reliability of two physicians who have reviewed a set of 100 pulmonary function tests:

Physician 1	**Physician 2**		
	Abnormal	Normal	
Abnormal	a 20	b 15	35
Normal	c 10	d 55	65
	30	70	

From the table above, we can begin to describe the agreement between physicians as:

$$observed\ agrement = \frac{a+d}{a+b+c+d} = \frac{75}{100} = 0.75\ or\ 75\%$$

However, 75% is an overestimate because some agreement will occur by chance. The index corrected for chance is called kappa (κ):

$$\kappa = \frac{observed\ agrement - chance\ agreemnt}{1 - chance\ agreement}$$

where

$$chance\ agreemnt = \frac{(a+b)(a+c)}{a+b+c+d} + \frac{(c+d)(b+d)}{a+b+c+d}$$

As a guide to interpretation you can use this table:

Value of Kappa	Strength of Agreement
< 0	Poor
0 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.0	Almost perfect

Although kappa is widely used in the literature, it has a limitation: As two observers classify an increasingly higher proportion of patients in one category or the other (such as normal or abnormal) the agreement by chance increases. Thus, kappa will increase even if the way raters interpret diagnostic tests does not change. One solution is to use another index called *phi.*

Phi

Phi (ϕ, rhymes with pie) is an index of agreement independent of chance. Thus, you will get similar values for phi whether the distribution of results is 50% positive and 50% negative or whether it is 90% positive and 10% negative, which is not true for kappa. Also, phi allows testing of significant differences between raters which kappa does not. Phi is calculated as follows (refer to table under kappa):

$$\Phi = \frac{ad - bc}{\sqrt{(a+b)(c+d)(a+c)(b+d)}}$$

Values for phi range from –1.0 (representing extreme disagreement) through 0.0 (chance agreement) to +1.0 (representing extreme agreement).

COMPARING A SINGLE SAMPLE WITH A POPULATION

Binomial Test

When comparing a proportion (or rate, or percentage) obtained from a single sample to a proportion from a known population, the problem is to decide if the sample proportion is unlikely to have occurred or not. As described in the chapter on basic statistics (hypothesis testing) we would get different outcomes every time we repeat the experiment. Some proportions may be higher, some lower, due to chance. Because our outcome variable is nominal, the possible proportions can only be whole numbers (2 out of 10 or 20%; we would never get a fractional outcome like 2.5 out of 10 or 20.5%). Because the outcomes are discrete numbers, the sampling distribution is discrete. For example, suppose you toss a coin ten times as an experiment. After ten tosses, you count the number of heads you got. If you repeat the experiment you will likely get a different number. If you repeated this experiment an infinite number of times and plotted the probability of each outcome, you would have a distribution as shown in Figure 11-4. This distribution is called a binomial distribution because the outcome can be only one of two possible types (or numbers; head or tails, one or zero, etc).

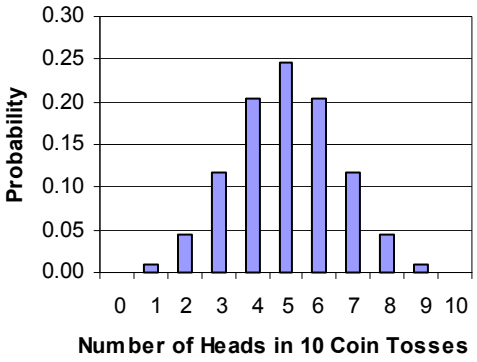


Figure 11-4. Binomial distribution showing the probability of getting various numbers of heads in 10 coin tosses (assuming the probability of getting heads on a single toss is 0.50).

Now, let's apply this distribution to a clinical situation. Suppose that you wanted to test the hypothesis that patients with acute respiratory distress syndrome (ARDS) have a different survival rate in your ICU than the national average, say 50%. The null hypothesis is that an observed proportion from a sample of patients is equal to the assumed population proportion of 0.50. You obtain a sample of 10 patients, 9 of whom survived.

As we discussed in the section on hypothesis testing, you need to decide if your observed proportion of 9/10 falls within the rejection regions on the distribution. In other words would sample proportions of equal or greater difference from 50% be unlikely (have probability less than α or 5%). We do not know beforehand if the sample proportion will be higher or lower than the population proportion, so we have a two-tailed rejection region. This example happens to correspond to Figure 11-4. Thus, we want to know the combined probability of observed patient proportions of 9/10, 10/10 on the right tail plus proportions or 1/10 and 0/10 in the left tail. If this total probability is less than 0.05, we reject the null hypothesis and conclude that our sample came from a population with a true probability of survival different from 50%.

The probability in the left tail is calculated using the Microsoft Excel spreadsheet equation:

=BINOMDIST(successes, sample size, hypothesized proportion, true)

This equation gives the cumulative probability for the distribution up to the proportion indicated by the number of successes in the sample (the value "true" tells the equation you want the cumulative probability). So, for our example we enter the equation

=BINOMDIST(1,10,0.5,true)

and receive the answer 0.011. The probability in the right tail is 1 minus the cumulative probability up to a proportion of 8/10. Thus we enter the equation

=1-BINOMDIST(8,10,0.5,true)

which gives the value of 0.011. Notice that because the distribution in this example is symmetrical, the area in the left tail is the same as the area in the right tail and we could have simply done one spreadsheet calculation and multiplied by 2. But the distribution is symmetrical only when the hypothesized proportion is 0.50.

The total area in the rejection regions is 0.011 + 0.011 = 0.022. This means that the probability of observing sample proportions as far away from the hypothesized population proportion as our sample was is about 2.2% Because we have set the significance level at $\alpha = 0.05$ or 5%, a *p* value from the hypothesis test ≤ 0.05 will indicate a significant difference. Therefore, we conclude that there is a significant difference between the survival rate in our ICU and that of the national average. In fact, we conclude that our survival rate is higher.

Z Test

The binomial distribution approaches the normal distribution as the sample size gets larger. When the sample size is large enough, we can use a z test with the normal distribution to simplify the calculations. The statistic is calculated as:

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

where

p = the sample proportion

p_0 = the hypothesized population proportion

n = the sample size

The sample size is considered large enough to use the normal approximation of the binomial distribution when

$$n \times p_0 > 5 \quad \text{and} \quad n \times (1 - p_0) > 5$$

Using our previous example, $n = 10$ and $p_0 = 0.50$ so $n \times p_0 = 5$ and $n \times (1 - p_0) = 5$, so we could not use the normal distribution. You can see that the larger n is, the more appropriate the approximation. Let's say our sample size was 20. Then

$$z = \frac{p - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.9 - 0.5}{\sqrt{\frac{0.5(0.5)}{20}}} = \frac{0.4}{\sqrt{0.0125}} = \frac{0.4}{0.11} = 3.58$$

The cut-off values of z for a two tailed test are $\pm$ 1.96. Because 3.58 is larger than +1.96, we reject the null hypothesis and conclude that a significant difference between the proportions is present. If the sample proportion had been 0.10, then z would be –3.58 which is less than –1.96 and again we would reject the null hypothesis.

The cutoff values for various significance levels (ie, different values of α) can be calculated with the Microsoft Excel spreadsheet equation

=NORMSINV(probability)

where probability is a for 1-α one tailed test and 1-α/2 for a two tailed test. For example, if the significance level is 0.05, and we want a two tailed test, 1-α/2 =1-0.025 = 0.975. Then NORMSINV(0.975) yields the value 1.96.

A 95% confidence interval (CI) for a sample proportion is constructed using:

CI = sample proportion $\pm$ 1.96 x standard error of proportion

$$CI = p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$$

If the sample proportion was 0.45 and the sample size was 30

$$CI = 0.45 \pm 1.96\sqrt{\frac{0.45(1-0.45)}{30}} = 0.45 \pm 1.96\sqrt{\frac{0.45 \times 0.55}{30}} = 0.45 \pm 0.18 = (27, 0.63)$$

which means we can be 95% confident that the interval between 0.27 and 0.63 contains the true proportion of patients who survive ARDS in our ICU.

COMPARING TWO SAMPLES, UNMATCHED DATA

Unpaired data means that there are data from two independent groups of experimental objects, such as two different groups of patients. For example, instead of comparing the survival rate of ARDS patients in our ICU with a benchmark value, suppose we wanted to compare survival rates in two ICUs. From now on, hypothesis tests will be more complicated so we will use statistical software instead of manual calculations. For most examples, we will be using SigmaStat by Jandel Scientific Software Corp.

Fisher Exact Test

The Fisher Exact Test is used for 2 x 2 contingency tables (exactly 2 rows and 2 columns).

Example. The contingency table for comparing survival rates in two ICUs might look like this:

Outcome	ICU A	ICU B
Lived	10	13
Died	2	4

Null Hypothesis: There is no significant difference between the proportion of patients that lived (or died) in ICU A compared to ICU B.

Most statistical programs will let you enter the data in columns and rows either as tabulated data (a contingency table like above) or as raw data like this:

ICU A	lived
ICU A	died
ICU B	lived
ICU B	died
ICU A	died
etc.	etc.

This latter format might be more convenient if you are transferring data from a spreadsheet with the actual experimental results. If we enter this example, we get a *p* value and possibly an interpretation as follows:

Report from statistics program:

The proportion of observations in the different categories, which define the contingency table, is not significantly different than is expected from random occurrence ($p = 1.000$). Thus, we conclude that the survival rates in the two ICUs are not different.

COMPARING TWO OR MORE SAMPLES, MATCHED DATA

Paired data result when patients act as their own controls or when two groups of patients are carefully matched for confounding factors like age, weight, sex, race, etc. When you are experimenting on pieces of identical equipment, obviously they are very well matched.

McNemar's Test

McNemar's test is an analysis of contingency tables that have repeated observations of the same individuals. The test is appropriate for

- determining whether or not individuals responded to treatment
- comparing results of two different treatments on the same people

Example. Suppose we wanted to test the acceptance of a new airway clearance technique among patients at your hospital with cyctic fibrosis. We record the patients' impressions before and after trying the new treatment. The raw data are:

Before Treatment	After Treatment
approve	approve
approve	approve
approve	approve
approve	disapprove
disapprove	disapprove
disapprove	disapprove
disapprove	disapprove
disapprove	disapprove
disapprove	disapprove
disapprove	don't know

Before Treatment	After Treatment
don't know	approve
don't know	disapprove
don't know	don't know

These data can be organized in the following contingency table;

		After Treatment	
Before Treatment	Approve	Disapprove	Don't Know
Approve	3	1	0
Disapprove	0	5	1
Don't Know	1	1	1

A similar table would have been created if we had compared one form of airway clearance to another, with all patients getting both treatments (acting as their own controls). Remember that the McNemar's test requires that the contingency table have exactly the same number of rows as columns (such as 2 x 2, 3 x 3, etc.)

Null Hypothesis: There is no significant difference between the paired proportions before and after treatment

Report from statistics program:

The proportion of observations in the different categories that define the contingency table is not significantly different than is expected from random occurrence ($p = 0.572$).

Thus, we conclude that patients' opinions did not change after trying the new treatment. This is not too surprising. Looking at the contingency table, we see that almost 70% of the patients (9/13) did not change their opinion after trying the treatment (3 approved, 5 disapproved, and 1 did not know).

COMPARING THREE OR MORE SAMPLES, UNMATCHED DATA

Chi-Squared Test

The Chi-Squared test can be used for analyzing contingency tables that are larger than 2 x 2.

Example. Suppose we wanted to test the effectiveness of three different drugs (or three dosages of one drug). The contingency table might be:

Outcome	Drug A	Drug B	Drug C
Effective	11	6	3
Not Effective	2	6	7

Null Hypothesis: The proportions (effective/total or not effective/total) are all equal (or equivalently, there is no association between drug and effectiveness).

Report from statistics program:

Power of performed test with alpha = 0.05: 0.671

The power of the performed test (0.671) is below the desired power of 0.800.

You should interpret the negative findings cautiously.

The proportions of observations in different columns of the contingency table vary from row to row. The two characteristics that define the contingency table are significantly related. ($p = 0.026$).

The two characteristics that define the contingency table are Outcome and Drug Type. Thus, we reject the hypothesis that the drugs had equal effectiveness. The data suggest that there is an association between the type of drug and effectiveness.

QUESTIONS

Definitions

Explain the meaning of the following terms:

- Contingency table
- Proportion
- Percentage
- Ratio
- Odds
- Rate
- Sensitivity
- Specificity
- Positive predictive ability
- Negative predictive ability
- Receiver Operating Characteristic (ROC) Curve

True or False

1. The kappa and phi statistics are used to evaluate the strength of agreement between two sets of data, such as the diagnoses of two physicians.
2. The choice of the cut-off value has no effect on a diagnostic test's sensitivity or specificity.
3. When two diagnostic tests are compared with ROC curves, the one with the most area under its curve is the most accurate test.

Multiple Choice

1. Which test is most appropriate for comparing the percentage of accidental extubations in your hospital with that of a known benchmark percentage?
    a. binomial test
    b. Fisher Exact test
    c. McNemar's test
    d. Chi-Square test
2. Suppose you were comparing the percentage of patients who changed their opinions of a treatment before and after trying it. What would be the best test?
    a. binomial test

b. Fisher Exact test

c. McNemar's test

d. Chi-Square test

3. If you wanted to compare the survival rates of three ICUs, what would be the most appropriate test?

   a. binomial test

   b. Fisher Exact test

   c. McNemar's test

   d. Chi-Square test

4. You want to compare the percentage of positive responses to therapy in two different groups of patients. What test is most appropriate?

   a. binomial test

   b. Fisher Exact test

   c. McNemar's test

   d. Chi-Square test

Chapter 12. Statistics for Ordinal Measures

Data on the ordinal level of measurement consist of discrete categories that have a particular order to them. Numbers are used to indicate relative high and low values of a variable (eg, a Likert-type scale). Like nominal measurements, ordinal measurements are often summarized in terms of percentages, proportions, ratios, and rates. When continuous measurements have extreme values or outliers, converting the data to the ordinal level (by sorting individual values by rank) often makes inferential tests more reliable.

DESCRIBING THE DATA

Contingency Table. Usually, the first step in describing the data is to create a contingency table. Such a table is used to display counts or frequencies of two or more ordinal variables. The contingency table below was used in a study of retinopathy in premature infants. The number and the percentage of patients can be given for each of the five stages of retinopathy:

Retinopathy Stage	Patients	Percent
0	6	30
I	6	30
II	2	10
III	3	15
IV	3	15
Total	20	100

Calculation of a mean or average is not appropriate for ordinal data. These data are often displayed with bar graphs or pie charts.

CORRELATION

Because ordinal data can be ranked according to their relative values, we can compare two sets of ordinal data to see if an increase (or decrease) in one variable corresponds to an increase (or decrease) in the other.

Spearman Rank Order Correlation

The Spearman Rank Order Correlation coefficient (also called Spearman's rho or rank correlation) is a nonparametric test that does not require the data points to be linearly related with a normal distribution about the regression line with constant variance. This statistic does not require that variables be assigned as independent and dependent.

The value of rho can take values from –1 through 0 to +1. A value of –1 indicates that high ranks of one variable occur with low ranks of the other variable. A value of 0 indicates there is no correlation between variables. A value of +1 indicates that high ranks of one variable occur with high ranks of the other variable.

Use this procedure when

- you want to measure the strength of association between pairs of ordinal variables or between an ordinal variable and a continuous variable;
- you want to reduce the effect of extreme values on the correlation of data on the continuous level of measurement. The Pierson Product Moment Correlation Coefficient is markedly influenced by extreme values and does not provide a good description of data that are skewed or that contain outliers. The solution is to transform the data into ranks (ie, convert from continuous to ordinal scale) then calculate the correlation.

Example: You suspect that Apgar scores are correlated with infants' birth weight. You collect the following data on a group of newborns:

Weight (grams)	Weight (rank)	APGAR Score
1,023	3	6
850	1	3
1,540	5	10
1,150	4	8
900	2	5

Null Hypothesis: Weight (when converted to an ordinal scale) is not associated with Apgar score.

Report from statistics program:

Correlation Coefficient = 0.900; $p = 0.0833$

Pairs of variables with positive correlation coefficients and p values below 0.050 tend to increase together. For pairs with negative correlation coefficients and p values below 0.050, one variable tends to decrease while the other increases. For pairs with p values greater than 0.050, no significant relationship exists between the two variables.

We must conclude that for this set of data, weight and Apagr score are not correlated. Even though the correlation coefficient (0.9) appears quite high, the p value tells us that it can be expected to occur by chance at least 8 out of 100 times. However, the p value of 0.0833 is not too much above the significance cut-off value of $p \leq 0.05$. This result should tell us that our experiment was a good pilot study. We are justified in collecting a larger sample that might yield positive results.

COMPARING TWO SAMPLES, UNMATCHED DATA

Mann-Whitney Rank Sum Test

This test is also called the Mann-Whitney U Test or the Wilcoxon Rank Sum Test. It is the nonparametric alternative to the unpaired *t* test. It tests the hypothesis that the medians of two different samples are the same.

Example: Your department has just implemented a Respiratory Therapy Consult Service. Using standardized assessment procedures, therapists assign triage scores to patients according to disease type and severity. You want to test the performance of a new supervisor (A) by comparing her assessment skills with a more experienced supervisor (B). Each supervisor assesses the same 8 patients on a scale of 1 to 5 with the following results:

Supervisor A	Supervisor B
1	1
3	4
2	3
4	4
5	5
2	3
4	4
2	2

Null Hypothesis: There is no difference between these groups of triage scores; the two sets of data were not drawn from populations with different medians.

Report from statistics program:

The difference in the median values of the two groups is not great enough to exclude the possibility that the difference is due to random sampling variability; there is not a statistically significant difference ($p = 0.442$).

Therefore, we conclude that the new supervisor's assessment skills are equivalent to the more experienced supervisor.

COMPARING TWO SAMPLES, MATCHED DATA

Wilcoxon Signed Rank Test

The Wilcoxon Signed Rank test is the nonparametric alternative to a paired *t* test. This test is appropriate when comparing treatments on the same individual or matched individuals.

Example: Your department has just completed a training program for therapists who will participate in a Respiratory Therapy Consult Service. The Consult Service employs standardized assessment procedures to assign triage scores to patients according to their disease type and severity. To test the effectiveness of the training program, you select a new therapist and ask him to assign triage scores (1 through 5) to a set of seven simulated patients before and after training.

Before	After
4	1
5	2
3	3
5	4
3	2
4	2
4	3

Null Hypothesis: There is no difference in the two groups of data.

Report from statistics program:

The change that occurred with the treatment is greater than would be expected by chance; there is a statistically significant difference ($p = 0.031$).

We conclude that the training program resulted in learning (assuming that the After scores were closer to the ideal scores for the simulated patients). In fact, we can see that before training, the therapist tended to score the patients higher (missing signs of illness severity) than after the training.

COMPARING THREE OR MORE SAMPLES, UNMATCHED DATA

Kruskall-Wallis ANOVA

This test is the nonparametric alternative to the one-way (or one factor) analysis of variance (ANOVA). You would use it when you want to see if the groups are affected by a single factor. It tests the hypothesis that the groups are the same versus the hypothesis that at least one of the groups is different from the others.

Example: Your department has been trying to decide on the appropriate length of a training program for therapists who will participate in a Respiratory Therapy Consult Service. The Consult Service employs standardized assessment procedures to assign triage scores to patients according to disease type and severity. To test the effectiveness of different training program lengths, you select three groups of seven new therapists and grade their assessment skills (grade 0 through 5) on simulated patients after each of three training sessions.

One Hour	Six Hours	12 Hours
0	1	4
5	4	5
2	3	3
3	4	4
2	5	5
4	3	3
3	4	4

Null Hypothesis: All three sets of data came from the same population; there are no differences among the groups.

Report from statistics program:

The differences in the median values among the treatment groups are not great enough to exclude the possibility that the difference is due to random sampling variability; the difference is not statistically significant ($p = 0.211$).

You conclude that training therapists longer than one hour is a waste of time.

Note that if the p value had been less than 0.05, we would be able to do pair-wise comparisons among the three groups to see which were different (see *Report of statistics program* in the next section).

COMPARING THREE OR MORE SAMPLES, MATCHED DATA

Friedman Repeated Measures ANOVA

This test is the nonparametric alternative to a one-way repeated measures analysis of variance (ANOVA). Use this test when you want to see if a single group of individuals was affected by a series of three or more different experimental treatments, where each individual received all treatments.

Example: You would like to see if the brand of incentive spirometer has any affect on patient compliance. You select a group of six post-operative surgical patients and have them try each of three different incentive spirometers. The patients rank their opinion numerically (0 = don't like, 1 = neutral, 2 = like). The data are entered into a table like this:

Spirometer A	Spirometer B	Spirometer C
0	1	2
1	1	1
0	2	2
0	1	2
1	0	2
1	1	2

Null Hypothesis: All three sets of data came from the same population; there are no differences among the groups.

Report from statistics program:

The differences in the median values among the treatment groups are greater than would be expected by chance; there is a statistically significant difference ($p = 0.027$)

To isolate the group or groups that differ from the others use a multiple comparison procedure.

All Pairwise Multiple Comparison Procedures (Student-Newman-Keuls Method) :

Comparison	**p<0.05**
C vs A	Yes
C vs B	Yes
B vs A	No

We conclude that the spirometers are not all the same in terms of patient preference. To find how they rank, the statistics program performed pairwise comparisons among the three spirometers. From that, we see that patients prefer spirometer C over A and C over B but they had equal preference for B compared to A.

QUESTIONS

Multiple Choice

1. If you wanted to test the hypothesis that birth weight is associated with Apgar score you would use which test?
    a. Spearman Rank Order Correlation
    b. Mann-Whitney Rank Sum Test
    c. Wilcoxon Signed Rank Test
    d. Kruskall-Wallis ANOVA
    e. Firedman Repeated measures ANOVA

5. Suppose you want to compare the assessment skills of two supervisors, each using assigning triage scores for a set of simulated patients. What test would you use?
    a. Spearman Rank Order Correlation
    b. Mann-Whitney Rank Sum Test
    c. Wilcoxon Signed Rank Test
    d. Kruskall-Wallis ANOVA
    e. Firedman Repeated measures ANOVA

6. Suppose you wanted to determine the improvement in assessment skill for a single person after a training session. What test would you use to see if their set of triage scores for simulated patients was different pre- and post-training?
    a. Spearman Rank Order Correlation
    b. Mann-Whitney Rank Sum Test
    c. Wilcoxon Signed Rank Test
    d. Kruskall-Wallis ANOVA
    e. Firedman Repeated measures ANOVA

7. You want to determine the most appropriate length of treatment based on a post-treatment assessment score. Three groups of patients are treated for 10, 20, and 30 minutes respectively. What procedure would you use to test the hypothesis that there was no difference in assessment scores among the three groups?
    a. Spearman Rank Order Correlation
    b. Mann-Whitney Rank Sum Test
    c. Wilcoxon Signed Rank Test
    d. Kruskall-Wallis ANOVA
    e. Firedman Repeated measures ANOVA

f. Kruskall-Wallis ANOVA

8. You would like to see if the brand of incentive spirometer has any affect on patient compliance. You select a group of 6 post-operative surgical patients and have them try each of 3 different incentive spirometers. The patients rank their opinion numerically (0 = don't like, 1 = neutral, 2 = like). What procedure would you use to test the hypothesis that there was no difference in opinion scores among the three brands?

   a. Spearman Rank Order Correlation

   b. Mann-Whitney Rank Sum Test

   c. Wilcoxon Signed Rank Test

   d. Kruskall-Wallis ANOVA

   e. Firedman Repeated measures ANOVA

Chapter 13. Statistics for Continuous Measures

Data on the continuous level of measurement can take on any value and include measurements on the interval and ratio levels (see chapter on basic statistics). Most of the data we are familiar with are measured at the continuous level (such as, pressure, volume, flow, weight, drug dosages, lab values, etc). Compared with nominal and ordinal levels, data at the continuous level contain more information and allow the widest range of statistical procedures.

TESTING FOR NORMALITY

A key assumption of the tests in this section is that the sample data come from a population that is normally distributed. If the sample data are not normally distributed, you will have to use one of the nonparametric tests described for data measured at the nominal or ordinal level.

Kolmogorov-Smirnov Test

The Kolmogorov-Smirnov procedure tests whether the distribution of a continuous variable is the same for two groups. That is, it tests the null hypothesis that two distributions are the same under the assumption that the observations from the two distributions are independent of each other. It is calculated by comparing the two distributions at a number of points and then considering the maximum difference between the two distributions. This test is heavily influenced by the maximum value in a set of numbers and should be used with caution if outliers are suspected.

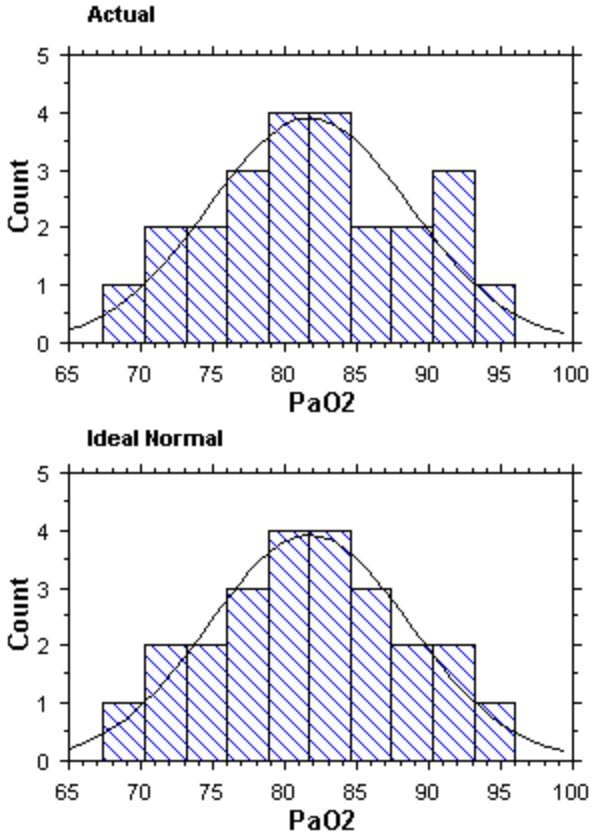
Example: You want to test the effect of prone positioning on the PaO_2 of patients in the ICU. Data are collected before and after positioning the patient and you intend to use a paired *t*-test. However, the first step should be to make sure the PaO_2 data are normally distributed. A statistics program such as SigmaStat (Jandel Scientific Software Inc.) will perform this test automatically. A statistics program called StatView (SAS Institute Inc.) allows you to enter your data in one column and the program then generates a column next to it that contains values from a normal distribution with the same mean and standard deviation as the variables in your sample data.

Null hypothesis: The two distributions are the same.

Report from statistics program:

StatView draws a histogram of the data superimposed on a normal curve for visual comparison (Figure 13-1). If the result of the K-S test is significant (ie, $p < 0.05$), then the Actual and Ideal variables are probably not from the same distribution. A significant difference implies that the Actual variable is not normally distributed, because the Ideal variable is. In the example shown in Figure 13-1, the *p* value is > 0.999, which is > 0.05 so we do not reject the null hypothesis. We assume the PaO_2 values are normally distributed and can use the *t* test.

Figure 13-1. Kolmogorov-Smirnoff test of normality. The top graph shows the actual sample data; the bottom graph shows ideal normal values from a normal distribution with the same mean and standard deviation as the sample. The normal curve is superimposed for visual comparison. The *p* value is > 0.999 which leads us to believe that there is no significant difference between the distributions. Therefore, the sample data are normally distributed.



TESTING FOR EQUAL VARIANCES

Another key assumption of the tests in this section is that the data in two or more samples have equal variances. If the data do not meet this assumption, you will have to use one of the nonparametric tests described for data measured at the nominal or ordinal level.

F Ratio Test

A comparison of the variances of groups of measurements can be useful to validate the assumptions of *t* tests and for other purposes. The *F* test is calculated as the ratio of two sample variances and shows whether the variance of one group is smaller, larger, or equal to the variance of the other group.

Example: One of the pulmonary physicians in your hospital has questioned the accuracy of your lab's PFT results. Specifically, he questions whether the new lab technician you have hired (Tech A) can produce as consistent results as the other technician who has years of experience (Tech B). You gather two sets of FEV_1 measurements on the same patient by the two techs. Since you are using only one patient, most of the variance in measurements will be due to the two technicians. You do not know what

the "true" FEV_1 is so you cannot determine which technician produces the most "accurate" results. You could determine an agreement interval to see how far apart individual measurements by the two technicians might be (see Error Intervals in Chapter 10). However, you are really more concerned about how much the two technicans' results vary in general. Because variance (the average squared deviation from the mean) represents random measurement error, you use the *F* test to compare the ability of the two techs to produce the same level of measurement accuracy. You enter the data into the statistics program as two columns of FEV_1 measurements:

Tech A	Tech B
1.06	0.98
1.23	1.24
0.98	0.91
1.29	1.26
etc.	etc.

Null Hypothesis: The variances of the two groups of data are the same.

Report from the statistics program:

	Var. Ratio	*F*-value	*p* value	95% Lower	95% Upper
Tech A/Tech B	0.232	0.232	0.0002	0.110	0.488

	Count	mean	variance	Std. Dev.	Std. Error
Tech A	30	1.206	0.143	0.378	0.069
Tech B	30	1.173	0.033	0.182	0.033

The *p* value is less than 0.05, so we reject the null hypothesis and conclude that the variances are not equal. From the table we can see that the variance of measurements made by Tech A (0.143) is much greater than the variance of measurements made by Tech B (0.033). We conclude that there may be something about the new technician's performance that is resulting in less consistent results than the more experienced technician.

CORRELATION AND REGRESSION

The concept of correlation implies that two variables covary; that is, a change in variable *x* is associated with change in variable *y*. Another basic assumption of this section is that the association between the two variables is linear. Once we have established that there is an association, we can use a given value

of one to predict the associated value of the other. For example, we can predict the normal breathing frequency based on a person's age. This type of prediction can be extended to more than two variables, such as the prediction of pulmonary function values based on height, weight, and age. Correlation and regression are also described in the chapter on basic statistics.

Pearson Product Moment Correlation Coefficient

The most common correlation coefficient with a continuous variable measurable on an interval level is the Pearson product moment correlation coefficient (Pearson *r*). The Pearson *r* statistic ranges in value from –1.0 (perfect negative correlation) through 0 (no correlation) to +1.0 (perfect positive correlation).

Example: You decide to evaluate a device, called the EzPAP, for lung expansion therapy. It generates a continuous airway pressure proportional to the flow introduced at its inlet port. However, the user's manual does not say how the set flow is related to resultant airway pressure, and we generally use the device without a pressure gauge attached. You connect a flowmeter and pressure gauge to the device and record the pressures (in cm H_2O) as you adjust the gas flow (in L/min) over a wide range. You enter the data into a statistics program and calculate the Pearson *r:*

Flow	Pressure
3	5
4	5
5	6
6	7
etc.	etc.

Null hypothesis: The two variables have no significant linear association.

Report from statistics program:

Correlation coefficient: 0.942

The *p* value is less than 0.001. The *p* value is from a test of the null hypothesis to see if the correlation coefficient is significantly different from 0. The *p* value is the probability of being wrong in concluding that there is a true association between the variables.

Pairs of variables with positive correlation coefficients and *p* values below 0.050 tend to increase together. For the pairs with negative correlation coefficients and *p* values below 0.050, one variable tends to decrease while the other increases. For pairs with *p* values greater than 0.050, is no significant relationship exists between the two variables.

We conclude that a high degree of linear correlation exists between pressure and flow. That will allow us to predict how much flow we must set to get a desired level of airway pressure. To do that, we perform a simple linear regression.

Simple Linear Regression

A simple regression uses the values of one independent variable (x) to predict the value of a dependent variable (y). Regression analysis fits a straight line to a plot of the data. The equation for the line has the form:

$$y = b_0 + b_1 x$$

where

y = the dependent variable,

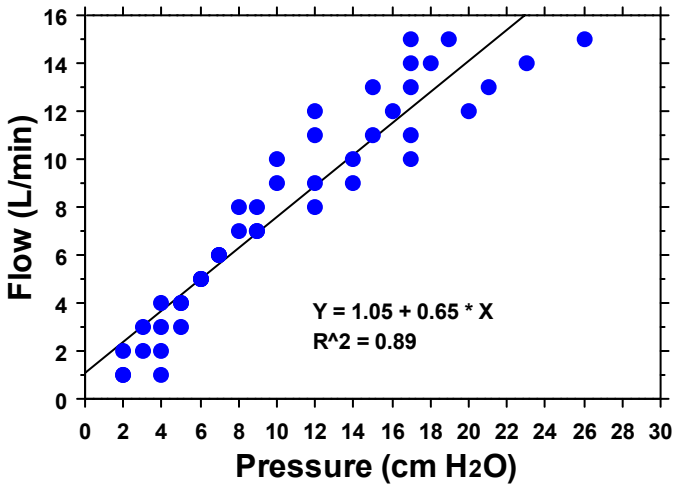
x = the independent variable

b_0 = the y intercept (point where the line crosses the y axis on the plot)

b_1 = the slope of the line ($\Delta y/\Delta x$)

Example: Using the data table from the EzPAP experiment above, we perform a simple linear regression. The purpose of the experiment was to allow us to predict the amount of flow required for a desired level of pressure. Therefore, we designate pressure as the (known) independent variable and flow as the dependent variable. In other words, the amount of flow we will set on the flowmeter connected to the EzPAP will depend on how much airway pressure we want the patient to receive.

Report from statistics program:



Normality Test: Passed ($p = 0.511$)

Constant Variance Test: Passed ($p = 0.214$)

	Coefficient	***p***
Y-Intercept	1.045	0.021
Pressure	0.653	<0.001

R squared: 0.89

Standard Error of Estimate = 1.482

The statistics program tested the two underlying hypotheses (a) that the data are normally distributed and (b) the two samples have equal variances. Both tests passed as indicated. The *p* values for the y-intercept and the slope indicate that they are both significantly different from zero. In other words, a significant correlation exists between pressure and flow.

The R^2 value (symbolized as R^2 in the graph above) is called the coefficient of determination (the square of the correlation coefficient, see chapter on basic statistics). The value of 0.89 means that 89% of the variation in flow is due to the variation in pressure, as we would guess from the obvious linear association and the high correlation coefficient. Only 11% of the variation in flow is due to random measurement error or nonlinearity in the relationship. Consequently, we can have a high degree of confidence in the accuracy of our predictions.

The differences between the measured values of *y* (in this case, flow) and those predicted by the regression equation are called *residuals*. Because the value predicted by the regression equation is the estimated mean value of repeated measurements and hence the best estimate of the true value of *x*, the residuals represent the random errors of measurement. The standard deviation of the residuals is called the *standard error of the estimate.* It is an estimate of the standard deviation of repeated measurements of *y* at any specific value of *x* and is thus an estimate of the imprecision of the measured values.

The regression equation says that the required flow (*y*) can be predicted by multiplying the desired pressure (*x*) by 0.65 (liters per minute per centimeter of water pressure) and adding 1.05 (liters per minute). To simplify the calculation, you round the numbers and inform the other therapists that they can get the required flow by multiplying the desired pressure by 0.7 and adding 1.

Multiple Linear Regression

Simple linear regression can be extended to cases where more than one independent (predictor) variable is present. When the independent variables are varied, they produce a corresponding value for the dependent (response) variable. If you are not sure if all independent variables should be used in the model, use *stepwise multiple linear regression* to identify the important independent variables from the set of possible variables. If the relationship does not fit a straight line or plane, use *polynomial* or *nonlinear* regression.

Multiple linear regression assumes an association between one dependent variable and an arbitrary number (symbolized by *k*) of independent variables. The general equation is:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_kx_k$$

where *y* is the dependent variable, $x_1, x_2, x_3, ...x_k$ are the *k* independent variables, and $b_0, b_1, b_2, ...b_k$ are the *k* regression coefficients. As the values of *x* vary, the value of *y* either increases or decreases depending on the sign of the associated regression coefficient *b.* An example of this would be the use of an equation to predict a pulmonary function value, like functional residual capacity, based on the patient's height and age.

Logistic Regression

Logistic regression is designed for predicting a qualitative dependent variable, such as presence or absence of disease, from observations of one or more independent variables. The qualitative dependent variable must be nominal and dichotomous (take only two possible values such as lived or died, presence or absence, etc.), represented by values of 0 and 1. The independent variables can be continuous, ordinal, or nominal. Independent ordinal variables can have names or be coded as 0 =

absence, 1 = level 1, 2 = level 2, etc. (this coding is helpful when calculating *P*, see below). Like linear regression, logistic regression can be simple (one independent variable) or multiple (many independent variables). The general logistic regression model is as follows:

$$P = \frac{e^Y}{1+e^Y}$$

where

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + ... + b_kx_k$$

P is interpreted to mean the probability of the dependent event happening given the specific values of the independent variables (not to be confused with the *p* from hypothesis testing). The predictors, $x_1, x_2, x_3, ...x_k$ are the *k* independent variables, $b_0, b_1, b_2, ...b_k$ are the *k* regression coefficients and *e* is the base of the natural logarithms (approximately equal to 2.718).

We can use the regression coefficients in two ways. One way is to use them with specific values for the independent variables in the above equation to calculate a value of *P*. The other way is to use the coefficients to calculate odds ratios. When *P* is small for all values of *x*, the *odds ratio* is approximately equal to the *relative risk.* For example, the relative risk of 10 in a smoking/lung cancer study would indicate that subjects who smoke are 10 times more likely to develop lung cancer than subjects who do not smoke.

Example: You conduct a study comparing two types of continuous positive airway pressure (CPAP) in premature infants. The dependent variable is the occurrence of complications such as nasal septum breakdown, deformation of the nares, or bleeding. The independent variables you think may influence the occurrence of complications include the type of CPAP delivery system, the CPAP fixation cap size, the duration of treatment, the infant's gestational age, and its birth weight. You enter the data into the statistics program like this:

Complication Present	**CPAP Type**	**Wrong Cap**	**Duration** (Days)	**Age** (Wks)	**Wt.** (gms)
0	conv	yes	0.84	28	1084
0	fluidic	yes	0.52	28	1287
1	conv	no	1.30	32	1749
0	conv	yes	5.20	28	1002
1	fluidic	no	5.40	28	713
etc.	etc.	etc.	etc.	etc.	etc.

where a value of 1 for Complication Present means that a complication was observed versus a value of 0 for no complications observed (remember the dependent variable must be dichotomous). The types of

CPAP compared were conventional and fluidic. The CPAP nasal prongs are secured by ties to a cotton cap on the infant's head, which come in several sizes; this variable could have been coded such that 1 means the wrong cap size was selected, 0 means the right size was selected. The wrong size cap leads to incorrect fit of the nasal prongs and may lead to complications. The Duration was the number of days of CPAP treatment. The Age is gestational age of the infant in weeks and Wt. is birth weight in grams. It is important to pay attention to how the data are named. Each statistics program has its own way of determining what level of nominal variable to use as the level against which other levels are compared. In our example, "complication present" was associated with fluidic CPAP compared to conventional because "conv" appeared first in an alphabetized list of CPAP types.

Report from statistics program:

	Coefficient	***p* Value**	**Odds Ratio**
CPAP Type (b_1)	1.265	0. 559	3.5
Wrong Hat (b_2)	-0.033	0.9448	1.0
Duration (b_3)	0.350	0.0023	1.4
Gest. Age (b_4)	-0.077	0.4459	0.9
Birth (b_2) Weight (b_5)	0.0002	0.6724	1.0
Constant (b_0)	-0.991		

The first column of the table gives the coefficients of the logistic regression. The second column shows the *p* values associated with the null hypothesis that each independent variable is unrelated to the incidence of complications. The *p* values indicate that only CPAP type and Duration are significantly related to complications. The Odds Ratio gives us the approximate increase in risk of using fluidic CPAP; about 3.5 times more likely to result in complications. Similarly, a unit increase in duration (1.0 day) leads to a 41% increase in risk of complications. The odds ratio for a unit increase in the independent variable is given by the equation:

$$odds\ ratio = e^{b}$$

where *b* is the coefficient of the independent variable of interest. Thus, if we want to calculate the odds ratio for two days (2 units) we would have

$$odds\ ratio = e^{2\times 0.35} = 2.718^{0.70} = 2.01$$

We can use the coefficients to calculate a specific probability of occurrence of complications. First, we calculate *y* using the equation given previously, the coefficients from the results table, and specific values of the independent variables. For example, suppose we want the probability of occurrence when CPAP type = fluidic (value of 1 versus 0 for conventional), wrong hat is yes (value of 1 versus 0 for right hat), duration = 2 day2, gestational age is 30 weeks and birth weight is 1,000 grams. The value of *y* is:

$$y = -0.991 + 1.265(1) - 0.033\ (1) + 0.350(2) - 0.077(30) + 0.0002(1000)$$

y = -1.17

Next, the value of *y* is substituted into the equation for *P:*

$$P = \frac{e^{Y}}{1+e^{Y}} = \frac{e-1.17}{1+e^{-1.17}} = \frac{0.31}{1.31} = 0.24$$

Thus, on the basis of this study, an infant with the listed characteristics would have a 24% chance of having complications due to CPAP using the fluidic system.

COMPARING ONE SAMPLE TO A KNOWN VALUE

One sample *t*-test

The one sample *t* test compares a sample mean to a hypothesized population mean and determines the probability that the observed difference between sample and hypothesized mean occurred by chance. The probability of chance occurrence is the *p* value. A *p* value close to 1.0 implies that the hypothesized and sample means are the same: The observed sample would probably not come from a population with the hypothesized mean. A small *p* value (less than 0.05) suggests that such a difference is unlikely (only one in 20) to occur by chance if the sample came from a population with the hypothesized mean. We would say that the sample mean is significantly different from the hypothesized value.

Example: You are considering the purchase of a new disposable ventilator circuit. The manufacturer claims that the tubing compliance is 1.5 cm H_2O. You take a sample circuit and connect it to a ventilator and lung simulator. A sample of 17 different tidal volume and airway pressure measurements yields a sample of 17 compliance values. You enter these values into a statistics program and request a one sample *t* test comparing your sample mean compliance to the manufacture's stated value of 1.5 cm H_2O.

Null Hypothesis: No difference exists between the sample mean and the hypothesized population mean for the tubing.

Report from statistics program:

Hypothesized Mean = 1.5

	Mean	*t*-value	*p* value	95% Lower	95% Upper
Compliance	1.653	1.062	0.304	1.348	1.958

The sample mean is 1.7 cm H_2O, which is greater than the hypothesized value of 1.5 cm H_2O. However, the *p* value is greater than 0.05 so we do not reject the null hypothesis. We conclude that the sample tubing compliance is not different from the manufacture's specification. The program gives the 95% confidence interval for the true mean value as being between 1.348 and 1.958 cm H_2O.

COMPARING TWO SAMPLES, UNMATCHED DATA

Unpaired *t*-test

The unpaired *t* test compares the means of two groups and determines the probability that the observed difference occurred by chance. The chance is reported as the *p* value. A *p* value close to 1.0 implies that the two sample means are the same, because the observed sample would probably not come from a

population with the hypothesized mean. A small *p* value (less than 0.05) suggests that such a difference is unlikely (only one in 20) to occur by chance if the sample came from a population with the hypothesized mean. We would say that the sample mean is significantly different from the hypothesized value. We could also say that the hypothesized difference between the means is zero.

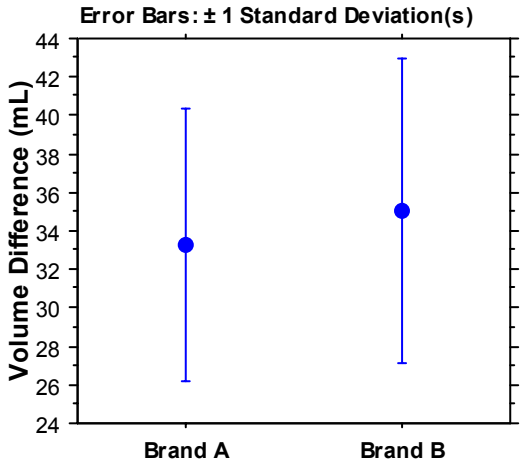
If your data fail the normality or equal variance tests, you should use the Mann-Whitney Rank Sum test.

Example: Suppose we want to change vendors for our disposable ventilator tubing. We want to be sure that the new brand (Brand A) has the same performance as the one we currently use (Brand B) in terms of volume lost due to compliance. We connect samples of both kinds of tubing to a ventilator and lung simulator and measure the lost volume over a wide range of set tidal volumes. We enter the two sets of lost volume data into the statistics program as follows:

Brand A	Brand B
33	28
30	25
25	27
etc.	etc.

Null Hypothesis: The average lost volume for Brand A is the same as Brand B.

Report from statistics program:



Normality Test: Passed ($p = 0.298$)

Equal Variance Test: Passed ($p = 0.096$)

Hypothesized Mean Difference = 0

Power of performed test with alpha = 0.050: 0.878

	Mean Diff.	*t* value	*p* value	95% Lower	95% Upper
Brand A – Brand B	-1.8	-0.684	0.4991	-7.0	3.5

	Count	mean	variance	Std. Dev.	Std. Error
Brand A	17	33.3	50.1	7.078	1.717
Brand B	17	35.1	49.3	7.021	1.702

The difference in the mean values of the two groups is not great enough to reject the possibility that the difference is due to random sampling variability. There is not a statistically significant difference between the input groups ($p = 0.4991$).

Power of performed test with alpha = 0.050: 0.070

The power of the performed test (0.070) is below the desired power of 0.800.

You should interpret the negative findings cautiously.

The results of two sample tests are often illustrated graphically as shown above. The dots in the graph represent the mean values for each sample. The the lines above and below the dots represent one standard deviation. These lines through the mean values are called error bars. Sometimes you see graphs where the error bars represent standard errors or possibly 95% confidence limits. Remember that error bars representing the *standard deviation* refer to the random error of individual measurements while bars that represent *standard error* refer to the random error of estimating the mean value.

The table shows that the mean lost volume with brand A is about 1.8 mL less than the volume lost with Brand B (because the mean difference is negative and was calculated as A minus B). However, this difference is due to chance and does not represent either a statistically significant or a clinically important difference ($p > 0.05$). We can be fairly confident that if we switch to Brand A, we will maintain the same standard of care. Although the power of the test was not very great, the consequences of making a Type II error are also not very great.

COMPARING TWO SAMPLES, MATCHED DATA

Paired *t*-test

The most common use of a paired *t* test is the comparison of two measurements from the same individual or experimental unit. The two measurements can be made at different times or under different conditions. The paired *t* test is used to evaluate the hypothesis that the mean of the differences between pairs of experimental units is equal to some hypothesized value, usually zero. A hypothesized value of zero is equivalent to the hypothesis that there is no difference between the two samples. The paired *t* test compares the two samples and determines the probability of the observed difference occurring by chance. The chance is reported as the *p* value. A small *p* value (less than 0.05) suggests that such a difference is unlikely (only one in 20) to occur by chance if the sample came from a population with the

hypothesized mean. We would say that the sample mean is significantly different from the hypothesized value. We could also say that the hypothesized difference between the means is zero.

The paired *t* test is more powerful than the unpaired *t* test, because it takes into account the fact that measurements from the same unit tend to be more similar than measurements from different units. For example, in a test administered before and after a therapeutic treatment, the unpaired *t* test may not detect small (but consistent) increases in each individual's outcome measurements. The paired *t* test is more sensitive to the fact that one measurement of each pair essentially serves as a control for the others.

A subject acting as his own control is called self pairing. The paired *t* test is also used for cases of natural pairing (such as twins or identical pieces of equipment) and artificial pairing (matching pairs of subjects on as many characteristics as possible). See the section on Matched Versus Unmatched Data in Chapter 10 on Basic Statistical Concepts.

If your data fail the normality or equal variance tests, you should use the Wilcoxon Signed Rank test.

Example: Two groups of asthmatic patients treated in the emergency department are entered into a study of aerosolized bronchodilators comparing the effects of standard racemic albuterol with levalbuterol. The two groups are matched for age, gender, race, and severity of illness. The outcome variable is length of stay (in hours) in the emergency department. The data are entered in the statistics program as

Racemic	Levalbuterol
33	28
30	25
25	27
etc.	etc.

Null Hypothesis: There is no difference in the mean length of stay between the two groups.

Report from statistics program:

The data are often graphed like that shown for the unpaired *t* test.

Normality test passed: ($p = 0.281$)

Hypothesized Mean Difference = 0

Power of performed test with alpha = 0.050: 0.050

	Mean Diff.	*t* value	*p* value	95% Lower	95% Upper
racemic – levalbut.	0.120	0.405	0.6.90	-0.501	0.741

	Count	Mean	Variance	Std. Dev.	Std. Error
racemic	20	2.3	1.073	1.036	0.232
levalbuterol	20	2.2	0.721	0.849	0.190

The change that occurred with the treatment is not great enough to exclude the possibility that the difference is due to chance (p = 0.690). The power of the performed test (0.050) is below the desired power of 0.800.

You should interpret the negative findings cautiously.

We conclude that this study should be considered an encouraging pilot study. A larger study will be needed to be confident that the length of stay in the emergency room is not different between the two drugs.

Comparing Three or More Samples, Unmatched Data

When a statistical test is performed at a given level of significance, such as 0.05, then the risk of a Type I error is controlled at the 5% level. However, when multiple tests are run at the 0.05 level, then the risk of a Type I error begins to increase. With c independent comparisons, the probability of getting at least one falsely significant comparison when there are actually no significant differences is given by:

$$probability\ of\ at\ least\ one\ false\ result = 1-(1-\alpha)^C$$

where α is the significance level. For example, suppose you want to know if patients treated with practice guidelines have better outcomes than those treated without guidelines. You want to compare two groups of patients on 3 variables; length of stay, number of complications, and patient satisfaction score. If you perform three t tests and at least one of them shows a significant difference, you will be tempted to conclude that a difference exists among the groups (especially if this conclusion agrees with your preconceived notions). However, if you report a significant difference and your significance level was 0.05, then you would be underestimating your risk of a Type I error. In fact, the real Type I error with three comparisons is $1 - (10 - 0.05)^3 = 0.14$, which is considerably higher than 0.05 that you assumed. Your real probability of making a Type I error in concluding there is a difference is 14% not 5%. This mistake appears in published studies surprisingly often.

The proper way to compare more than two mean values is to use the analysis of variance (ANOVA) procedure.

One Way ANOVA

A one way (or one factor) analysis of variance looks at values collected at one point in time for more than two different groups of subjects. It is used when you want to determine if the means of two or more different groups are affected by a single experimental factor. ANOVA tests the null hypothesis that the mean values of all the groups are the same versus the hypothesis that at least one of the mean values is different. If the test suggests that you reject the null hypothesis, other tests can be used to find which means are different. These are called post hoc (unplanned or after the fact) tests and are usually conducted as pairwise comparisons among all the possible pairs of groups (one common example is the Tukey Test).

If your data fail the normality or equal variance tests, you should use the Kruskal-Wallis ANOVA on Ranks.

Example: Suppose you want to compare the noise levels produced by four different brands of mechanical ventilator. You collect sound intensity (in decibels, dB) measurements from several ventilators in of each type. You enter the data into the statistics program as:

Bear 3	PB 840	Siemens 300	Versamed
58.18	50.68	51.83	57.72
58.53	49.95	52.13	56.46
57.36	50.21	51.56	56.82
58.68	49.61	52.83	56.98

Null Hypothesis: There are no differences in mean sound intensity between any two of the four groups of measurements.

Report from statistics program:

The data from this type of analysis are often illustrated with a graph similar to the one shown for the unpaird *t* test, except that more than two groups would be shown.

Normality test passed: ($p = 0.742$)

Equal variance test passed: ($p = 0.984$)

Power of performed test with alpha = 0.050: 1.000

	Mean	Std. Dev.	Std. Error
Bear 3	58.188	0.592	0.296
Versamed	56.994	0.531	0.265
Siemens 300	52.088	0.546	0.273
PB 840	50.111	0.452	0.226

The differences in the mean values among the treatment groups are greater than would be expected by chance. The p value is < 0.001, indicating a statistically significant difference.

Pair-wise multiple comparisons using the Tukey Test:

	Diff. of Means	$p < 0.05$
Bear 3 vs. PB 840	8.077	Yes
Bear 3 vs. Siemens 300	6.100	Yes
Bear 3 vs. Versamed	1.194	Yes
Versamed vs. PB 840	6.883	Yes
Versamed vs. Siemens 300	4.906	Yes
Siemens 300 vs. PB 840	1.977	Yes

We conclude that none of the ventilators produces the same noise level. They are ranked from loudest to quietest in the first table.

Two Way ANOVA

In a one way ANOVA we were comparing values *between groups.* In a two way (or two factor) ANOVA, we are comparing values *within groups as well as between groups.* This analysis would be appropriate for looking at comparisons of groups at different times as well as the differences within each group over the course of the study.

In a two factor ANOVA, there are two experimental *factors*, which are varied for each experimental group. The two or more different values of each factor are called *levels.* The test is for differences between samples grouped according to the levels of each factor and for *interactions* between the factors.

Data for two way ANOVA

		Factor 1
Factor 2	Level 1	Level 2
Level 1	Subject A Subject B Subject C	Subject H Subject I Subject J
Level 2	Subject D Subject E Subject F	Subject K Subject L Subject M

A two factor ANOVA tests three hypotheses:

1. There is no difference among the levels of the first factor;
2. There is no difference among the levels of the second factor;
3. There is no interaction between factors. That is, if there is any difference among levels of one factor, the differences are the same regardless of the second factor level.

Example: Suppose you want to test the effect of gender on ventilator weaning time (in hours) using two modes of ventilation, synchronized intermittent mandatory ventilation (SIMV), and pressure support (PS). One factor is gender, with two levels (male and female). The other factor is mode of ventilation, with two levels (SIMV and PS). A study including 12 patients might look as follows:

	Mode of Ventilation	
Gender	SIMV	PS
Male Weaning time (hours)	38 32 35	15 18 22
Female Weaning time (hours)	51 49 55	59 61 66

The data are entered into the statistics program like this:

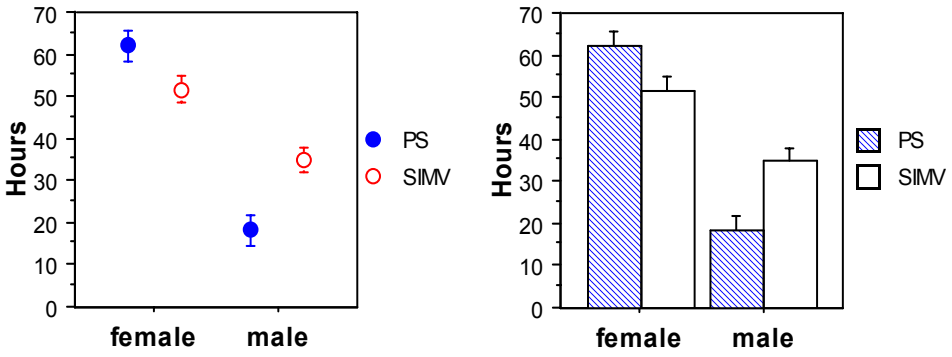
Gender	Mode	Hours
male	SIMV	38
male	SIMV	32
male	SIMV	35
male	PS	15
male	PS	18
male	PS	22
female	SIMV	51
female	SIMV	49
female	SIMV	55
female	PS	59
female	PS	61
female	PS	66

Null Hypohtheses:

1. There is no difference between SIMV and PS.
2. There is no difference between males and females.
3. There is no interaction between gender and mode.

Report from statistics program:

This type of data are often graphed using either points or bars. Make sure you specify what the error bars represent (either standard deviation or standard error):



Notice that the fill pattern for the dots or the bars is selected so that it will reproduce accurately on a black and white copier. Authors sometimes use colors only to distinguish between groups on a graph and this detail is lost when readers make photocopies.

Normality Test: Passed ($p = 0.435$)

Equal Variance Test: Passed ($p = 0.959$)

Power of performed test with alpha = 0.0500: for Gender : 1.000

Power of performed test with alpha = 0.0500: for Mode : 0.207

Power of performed test with alpha = 0.0500: for Gender x Mode : 1.000

	Mean	Std. Error
male	26.667	1.349
female	56.833	1.349

	Mean	Std. Error
SIMV	43.333	1.349
PS	40.167	1.349

	p value
Gender	<0.001
Mode	0.135
Gender x Mode	<0.001

The difference in the mean values between the two genders is greater than would be expected by chance after allowing for effects of differences in Mode. There is a statistically significant difference ($p < 0.001$). To isolate which group(s) differ from the others use a multiple comparison procedure.

The difference in the mean values between the modes is not great enough to exclude the possibility that the difference is just due to random sampling variability after allowing for the effects of differences in Gender. There is not a statistically significant difference ($p = 0.135$).

The effect of gender depends on which mode is present. There is a statistically significant interaction between gender and mode. ($p = <0.001$)

Pairwise multiple comparison procedure (Tukey Test):

	Mean	Std. Error
female x PS	62.000	1.908
female x SIMV	51.667	1.908
male x SIMV	35.000	1.908
male x PS	18.333	1.908

Based on these results, we conclude that males wean faster than females but there is no difference between the modes. Furthermore, females take longer when they are on PS and males take longer when they are on SIMV.

COMPARING THREE OR MORE SAMPLES, MATCHED DATA

One Way Repeated Measures ANOVA

A one way (or one factor) analysis of variance looks at values collected at more than one point in time for a single group of subjects. It is used when you want to determine if a single group of individuals was affected by a series of experimental treatments or conditions. ANOVA tests the null hypothesis that all the mean values of all the groups are the same versus the hypothesis that at least one of the mean values is different. If the test suggests that you reject the null hypothesis, other tests can be used to find which means are different. These are called post hoc (unplanned or after the fact) tests and are usually conducted as pairwise comparisons among all the possible pairs of groups (one common example is the Tukey Test).

If your data fail the normality or equal variance tests, you should use the Friedman Repeated Measures ANOVA on Ranks.

Example: You have a care path for pediatric asthmatic patients that requires therapists to perform assessments at specified intervals of 2, 3, and 4 hours from the start of treatments. One of these assessment is forced expiratory volume in one second (FEV_1) which requires both time and equipment to perform bedside spirometry. You suspect that FEV_1 does not change that much over 4 hours and may thus be eliminated as an assessment procedure. On a group of 10 patients treated with the care path, you record FEV_1 values after 2, 3, and 4 hours of bronchodilator treatment:

2 Hours	3 Hours	4 Hours
1.23	1.81	1.81
1.38	1.38	1.77
1.01	1.46	1.56
0.96	1.41	1.60
1.31	1.44	1.77
1.12	1.26	1.28
1.09	1.48	1.72
1.19	1.46	1.80
1.27	1.00	1.63
1.17	1.23	1.80

Null Hypothesis: there are no differences in mean FEV_1 between any two of the three groups of measurements.

Report from statistics program:

The data from this type of analysis are often illustrated with a graph similar to the one shown for the unpaird *t* test, except that there would be more than two groups shown.

Normality test passed: ($p = 0.405$)

Equal variance test passed: ($p = 0.204$)

Power of performed test with alpha = 0.050: 1.000

	Mean	Std. Dev.	Std. Error
2 Hours	1.173	0.132	0.0419
3 Hours	1.392	0.207	0.0656
4 Hours	1.676	0.166	0.0524

The differences in the mean values among the treatment groups are greater than would be expected by chance; the difference is statistically significant ($p < 0.001$).

Pairwise multiple comparisons using the Tukey Test:

	Difference of Means	$p < 0.05$
4 Hours vs. 2 Hours	0.503	Yes
4 Hours vs. 3 Hours	0.284	Yes
3 Hours vs. 2 Hours	0.219	Yes

We conclude FEV_1 does indeed increase at each assessment period. Although the differences between any two periods is statistically significant, we may conclude that there is not enough clinically important difference to justify the time and effort to collect the FEV_1 data.

Two Way Repeated Measures ANOVA

In a one way ANOVA we were comparing values *between groups.* In a two way (or two factor) ANOVA, we are comparing values *within groups as well as between groups.* This analysis would be appropriate for looking at comparisons of groups at different times as well as the differences within each group over the course of the study.

In a two factor ANOVA, there are two experimental *factors*, which are varied for each experimental group. Either or both of these factors are repeated treatments on the same group of individuals. The two or more different values of each factor are called *levels.* The test is for differences between the different levels of each factor and for *interactions* between the factors.

A two factor ANOVA tests three hypotheses:

1. There is no difference among the levels of the first factor;
2. There is no difference among the levels of the second factor;
3. There is no interaction between factors. That is, if there is any difference among levels of one factor, the differences are the same regardless of the second factor level.

Design for two way repeated measures ANOVA with one repeated measure (factor).

	Factor 1	
Factor 2	Level 1	Level 2
Level 1	Subject A Subject B Subject C	Subject A Subject B Subject C
Level 2	Subject D Subject E Subject F	Subject D Subject E Subject F

Data entry for two way repeated measures ANOVA with one repeated measure (factor). Note: Data entry format shown for SigmaStat, other programs require different formats.

Subject	Factor 1	Factor 2	Measurement
A	Level 1	Level 1	Value 1
A	Level 2	Level 1	Value 2
B	Level 1	Level 1	Value 3
B	Level 2	Level 1	Value 4
C	Level 1	Level 1	Value 5
C	Level 2	Level 1	Value 6
D	Level 1	Level 2	Value 7
D	Level 2	Level 2	Value 8
E	Level 1	Level 2	Value 9
E	Level 2	Level 2	Value 10
F	Level 1	Level 2	Value 11
F	Level 2	Level 2	Value 12

Design for two way repeated measures ANOVA with two repeated measures (factors).

		Factor 1	
Factor 2		Level 1	Level 2
Level 1	Subject A	Subject A	
	Subject B	Subject B	
	Subject C	Subject C	
Level 2	Subject A	Subject A	
	Subject B	Subject B	
	Subject C	Subject C	

Data entry for two way repeated measures ANOVA with two repeated measures (factors). Note: Data entry format shown for SigmaStat, other programs require different formats.

Subject	Factor 1	Factor 2	Measurement
A	Level 1	Level 1	Value 1
A	Level 2	Level 1	Value 2
B	Level 1	Level 1	Value 3
B	Level 2	Level 1	Value 4
C	Level 1	Level 1	Value 5
C	Level 2	Level 1	Value 6
A	Level 1	Level 2	Value 7
A	Level 2	Level 2	Value 8
B	Level 1	Level 2	Value 9
B	Level 2	Level 2	Value 10
C	Level 1	Level 2	Value 11
C	Level 2	Level 2	Value 12

Example: Suppose you want to assess the effects on oxygenation of a mode of mechanical ventilation called automatic tube compensation (ATC) using two different ventilators, the Dräger Evita 4 and the Puritan Bennett 840. The outcome variable is PaO_2. You select three patients on the Dräger and three on the Bennett 840. For each patient, you select two different levels of tube compensation. This experiment uses a repeated measures design with one factor (Tube Compensation) repeated.

	Tube Compensation Setting	
Ventilator	0%	100%
Dräger	100	125
	85	130
	90	105
PB 840	55	55
	75	80
	70	75

You enter the data into the statistics program like this:

Subject	**Tube Comp. (%)**	**Ventilator**	**PaO_2 (mm Hg)**
A	0	Dräger	90.8
A	100	Dräger	110.0
B	0	Dräger	84.2
B	100	Dräger	92.0
C	0	Dräger	79.6
C	100	Dräger	84.7
D	0	Puritan Bennett	87.9
D	100	Puritan Bennett	106.0
E	0	Puritan Bennett	77.5

Subject	Tube Comp. (%)	Ventilator	PaO_2 (mm Hg)
E	100	Puritan Bennett	91.4
F	0	Puritan Bennett	73.8
F	100	Puritan Bennett	90.4

Null Hypohtheses:

1. There is no difference between Dräger and Puritan Bennett.
2. There is no difference between 0% ATC and 100% ATC.
3. There is no interaction between ventilator and ATC level.

Report from statistics program:

This type of data are often graphed using either points or bars as shown for the two way ANOVA without repeated measures. Make sure you specify what the error bars represent (either standard deviation or standard error):

Normality Test: Passed (p = 0. 0.272)

Equal Variance Test: Passed (p = 0. 0.060)

Power of performed test with alpha = 0.0500: for Ventilator : 0.0503

Power of performed test with alpha = 0.0500: for ATC : 0.990

Power of performed test with alpha = 0.0500: for Ventilator x ATC : 0.0865

	Mean	Std. Error
Dräger	90.252	5.003
Puritan Bennett	87.845	5.003

	Mean	Std. Error
0% ATC	82.338	3.712
100% ATC	95.759	3.712

	p value
Ventilator	0.751
ATC Level	0.004
Ventilator x ATC	0.285

The difference in the mean values between the different ventilators is not great enough to exclude the possibility that the difference is just due to random sampling variability after allowing for the effects of differences in ATC. There is not a statistically significant difference (p = 0.751).

The difference in the mean values between the different levels of ATC is greater than would be expected by chance after allowing for effects of differences in type of ventilator. There is a statistically

significant difference (p = 0.004). To isolate which group(s) differ from the others use a multiple comparison procedure.

The effect of different ventilators does not depend on the level of ATC present. There is not a statistically significant interaction between ventilator and ATC. (p = 0.285)

Pairwise multiple comparison procedure (Tukey Test):

	Mean	Std. Error
Dräger x 0.000	84.928	5.249
Dräger x 100.000	95.576	5.249
Puritan Bennett x 0% ATC	79.748	5.249
Puritan Bennett x 100% ATC	95.942	5.249

Based on these results, we conclude that there is no difference in the way ACT works on the two ventilators in terms of oxygenation. However, using ATC does improve oxygenation.

QUESTIONS

Multiple Choice

1. One of the assumptions of the *t*-test is that the data you are analyzing are normally distributed. Before using this test, you should confirm this assumption using:
    a. Kolomogorov-Smirnov test
    b. F ratio test
    c. Pearson Product Moment Correlation Coefficient
    d. Regression analysis
2. You have a set of data showing patients' room air PaO_2 at sea level and at a simulated altitude of 8,000 feet. If you wanted to use this data to predict PaO_2 at altitude from PaO_2 at sea level, you would use:
    a. Kolomogorov-Smirnov test
    b. F ratio test
    c. Pearson Product Moment Correlation Coefficient
    d. Regression analysis
3. You have a new blood gas machine that seems to be producing inconsistent results. In order to test the hypothesis that the variance of this machine's measurements is different from your old machine, you would use:
    a. Kolomogorov-Smirnov test
    b. F ratio test

c. Pearson Product Moment Correlation Coefficient

d. Regression analysis

4. Suppose you want to determine whether or not training time is associated with competency scores. What would you use:

   a. Kolomogorov-Smirnov test

   b. F ratio test

   c. Pearson Product Moment Correlation Coefficient

   d. Regression analysis

5. The procedure that uses one independent variable to predict the value of one dependent variable is:

   a. Simple linear regression

   b. Multiple linear regression

   c. Logistic regression

6. If you have more than one independent (predictor) variables you would use:

   a. Simple linear regression

   b. Multiple linear regression

   c. Logistic regression

7. If you want to predict the presence or absence of a disease based on one or more independent variables, you must use:

   a. Simple linear regression

   b. Multiple linear regression

   c. Logistic regression

8. If we want to use the regression coefficients to calculate odds ratios, we need:

   a. Simple linear regression

   b. Multiple linear regression

   c. Logistic regression

9. A study suggests that the average length of stay reported by pediatric ICUs in the United States is 8.5 days. To test the hypothesis that your pediatric ICU has a shorter length of stay, you would use:

   a. One sample *t*-test

   b. Unpaired *t*-test

   c. Paired *t*-test

   d. ANOVA

10. You have been giving the same final exam year after year to a class of health care students. You have a set of 6 scores representing the average class performance. What would you use to test the hypothesis that the average class score has been decreasing over time?
    a. One sample *t*-test
    b. Unpaired *t*-test
    c. Paired *t*-test
    d. ANOVA
11. Two sets of patients are entered into a study. One set is treated with a new fast-acting bronchodilator and the other a placebo. What test would you use to make sure the two groups of patients had similar FEV_1 values at the start of the study?
    a. One sample *t*-test
    b. Unpaired *t*-test
    c. Paired *t*-test
    d. ANOVA
12. In the study above, what test would you use to compare the mean FEV_1 values before and after treatment?
    a. One sample *t*-test
    b. Unpaired *t*-test
    c. Paired *t*-test
    d. ANOVA

SECTION V PUBLISHING THE FINDINGS

Chapter 14. The Paper

Any research project that is important enough to require careful thought in design, and sustained effort in data gathering, merits a written report at its conclusion. Such a report serves at least two purposes. First, it forces the researcher to review the entire thought process involved, from the initial recognition of a clinical problem through the formation of a conceptual model relating the associated variables, to the comparison of this model with reality using actual measurements. Any flaws or ambiguities in this train of thought will become evident once the events are recreated in a written report. Someone once said that to understand a subject, one should teach it. A research report is a teaching instrument in the sense that it provides a concise review of the background and current perspective on a specific topic. A second, more obvious function of a research report is to expand the scientific community's knowledge base. If the report is published, it will become an immortal piece of the ever-changing puzzle we call scientific knowledge.

Publication of an article in a scientific journal can also result in personal benefits for the author. It is a measure of professional success and may contribute to tenure and status. In the extreme case, one may find oneself in a "publish or perish" situation if employed as a faculty member of a university.

In this chapter we will review the basic procedures for publishing a full-length manuscript. In the next two chapters we will discuss more specialized aspects of publication including how to submit an abstract and how to prepare a poster presentation based on the abstract.

SELECTING AN APPROPRIATE JOURNAL

In order for research to be accepted by the scientific community, it must "fit in." A journal with the appropriate readership must be selected. A topic that may seem original and interesting to one audience may be insulting to another. For this reason, the researcher should select the journal in which he or she hopes to publish the study findings before writing the paper. Several considerations in choosing an appropriate journal will now be reviewed.

Writing Style: By reading journals one will get a feel for the subject matter and level of complexity of articles selected by the editors. Some journals deal primarily with the more abstract topics of basic research (e.g., *The Journal of Applied Physiology),* whereas others emphasize more clinically relevant matters (e.g., *Critical Care Medicine* or *Respiratory Care).* Also, some journals presume a high level of mathematical sophistication on the part of the reader, as indicated by frequent articles containing calculus or complex statistical procedures. If an article is based on a rather simplistic algebraic model, it is likely to be rejected by such a journal.

Before beginning to write, look for an "Instruction for Authors" page included in many journals. This will provide detailed instructions concerning writing style, reference format, and preparation of illustrations. It will also indicate whether or not the journal charges a fee to publish the manuscript.

Types of Authors: Respiratory therapists sometimes have difficulty publishing a manuscript in a journal whose authors are almost exclusively MDs or PhDs. Review past issues to determine whether or not

professionals with your credentials have published in the journal in question. If an original investigation would be of interest mainly to respiratory therapists, it is more likely to be accepted by a journal in which other therapists contribute articles. When published in such a journal, your article will get the greatest exposure to the most appropriate audience.

Types of Readers: Health care professions have become highly diversified, with journals catering to each specialty and subspecialty. In addition to journals of interest to specific professionals, such as nurses or cardiopulmonary technologists, there are journals that are read by a variety of medical personnel (e.g., *Critical Care Medicine).* Selecting an appropriate journal will depend somewhat on the research topic's degree of specialization, as well as the intended audience.

Indexing: One of the factors that determines how widely read an article will be is how easily it can be found by other investigators. In searching the literature, other researchers routinely use printed bibliographies such as Index Medicus or CINAHL (Cumulative Index to Nursing and Allied Health Literature). Services such as Current Contents, or various computer databases, provide reviews of current published titles based on printed bibliographies. However, not all journals are covered by these sources. Select a journal that is indexed in a major reference source.

Peer Review. Most respectable scientific journals are peer-reviewed. Peer-review means that the editor will send copies of a prospective manuscript to consultants who are familiar with the subject area of the paper. A biostatistician may also be involved to review the study design and data analysis. These consultants will provide the editor and the author with detailed criticisms of the manuscript. Such criticisms are beneficial to the author, who obtains a better idea of contemporary professional standards; to the journal, by establishing credibility; and to the reader, who can be more confident of the validity of the report. Having a paper accepted by a peer-reviewed journal is more difficult than gaining acceptance from a so-called "throwaway" journal. The subscriptions to this latter type are generally offered free because they are primarily vehicles for the advertisements of medical equipment manufacturers. Obviously, an article published in a journal with strict editorial policies means more to the author and the reader than one published in a less sophisticated periodical.

GETTING STARTED

Authorship: All persons listed as authors must have participated in the reported work and in the shaping of the manuscript. All must have proofread the submitted manuscript. All should be able to publicly discuss and defend the paper’s content. Authorship is not based solely on solicitation of funding, collection or analysis of data, provision of advice, or similar services. Persons who provide such ancillary services may be recognized in an Acknowledgements section, but written permission is required from the person acknowledged. (see also the section called Who Should Write It? in the chapter on writing the case report)

The Rough Draft: Writing a high-quality manuscript is hard work, but the exertion of writing and rewriting a paper results in something that is worth reading. The beginner should be aware that a good paper is generally rewritten at least three to eight times before the final draft is completed. In light of this, it is helpful to have access to a word processor or at least an assistant to type each redaction so that enthusiasm for the project is not drained by the drudgery of retyping the text.

Always start with an outline of the paper. Never just start writing off the top of your head. Sometimes it is easier to start with the results because that is what is most fresh in your mind and it is fairly simple. That helps you get over any “writer’s block”. Next fill in the methods, making sure that they correspond with the results. Then write the discussion and conclusions. After that, you should be well prepared to

write an interesting, brief, and informative introduction, making sure you include the hypothesis. Finally, write the abstract. Then, leave the manuscript alone for a week or two. After a "cooling off" period, go back and proofread the manuscript. Now that the project is not so fresh in your mind you will be better able to read it with the eyes of your audience, who are totally new to it. This process will help you avoid reading between the lines instead of explaining things in enough detail so that someone else can follow your train of thought.

Copies of all reference articles should be available before you begin writing. The format for listing references in the text may vary with each journal. For example, some journals require references to be numbered consecutively with superscripts. However, for the first few drafts, simply list the last name of the article's first author and the year of publication. Each time a reference is added to the text, it is noted in complete bibliographical form, on a separate sheet. This format minimizes the work of changing the order and placement of references as the manuscript is revised. As the paper nears completion, the manner in which the references are listed can be changed to accommodate the style of a particular journal (e.g., consecutive numbering).

Before the writing begins, establish how the responsibility for writing will be allocated and in what order the authors will be listed. Actual writing of the manuscript usually proceeds most smoothly if one person is designated as the primary author. This person is responsible for writing the first draft. Contributing authors edit the first few drafts and provide input. The primary author then coordinates subsequent revisions and prepares the final manuscript for submission to a journal.

A primary author who has not published a manuscript before will find it extremely helpful to find an experienced colleague to act as a mentor. Ideally, you should consult someone who has knowledge of or interest in your research topic and who has published several papers. If one is fortunate enough to find such an advisor, consider offering to include his or her name as a coauthor. In this way both parties, and ultimately the manuscript, will benefit.

THE STRUCTURE OF A PAPER

There are five main sections to every scientific manuscript, excluding the title. We review each section in the order in which they appear, which is usually not the order in which they are written. Refer to the model manuscript in the Appendix for specific examples.

Title

There is a certain art to wording a title. It should not be too vague and all acronyms should be spelled out. For example, a poor title might be: An Investigation of CPAP. A better title would be: The Effects of Early Continuous Positive Airway Pressure on Length of Stay in the Neonatal Intensive Care Unit. The title should serve to introduce the topic and catch the reader's attention, while avoiding any sense of drama or marketing of conclusions. The best way to learn is to look at how titles are worded in respected medical journals.

Abstract

An abstract is required by many journals and usually appears after the title in the published article. As a practical matter, it is generally written after the main body of the manuscript is finished. The purpose of the abstract is to provide a short, concise summary of the study for the reader who may not have the time to read the whole article. The abstract is essentially a miniature, condensed version of the paper containing the same format: introduction, methods, results, and conclusion. However, the results and

methods are more important than the introduction and conclusion, and since space is limited, no more than one or two sentences should be devoted to a problem statement or discussion of conclusions.

In writing the abstract, avoid simply listing or describing the contents of the paper. The abstract should be informative and complete as a stand-alone piece of communication. Avoid phrases such as "Results will be given…" which force the reader to consult the text and defeat the purpose of a summary. On the other hand, the abstract should not contain information that is not presented in the text. Do not include opinions or conclusions without providing the evidence on which they were based.

Most journals specify a maximum length for the abstract. Requirements vary, but an appropriate length is approximately 250 words or less. In striving for brevity, remember to provide the same quality of grammar as the main body of the article. Do not eliminate essential prepositions, adjectives, and conjunctions such as "of," "the," and "a," but do try to eliminate words and "flowery language" that add nothing to the informative content of the summary.

Introduction

The introduction consists of two or three paragraphs that explain the purpose of the paper. It should contain a brief description of the background and significance of the research topic and include a few key references. The introduction should provide the reader with an awareness of the context and scope of the study.

A statement of the research problem or hypothesis should be included. Not including the hypothesis is a common mistake among beginners. Describing the hypothesis or research problem in the introduction sets the stage for the methods (which must describe how the hypothesis was tested), the results (which must correspond to all the methods described), and the discussion (which tells how the results addressed the hypothesis).

The references cited in the introduction should support the theoretical framework of the hypothesis, although an in-depth explanation should be saved for the discussion section. The introduction should also contain definitions of the general concepts treated in the manuscript. Frequently used terms can be abbreviated after first being spelled out in the opening paragraphs.

Methods

The purpose of this section is to explain to the reader exactly what was done to answer the research question and/or test the hypotheses described in the introduction. It should be brief, but must contain enough information to allow other researchers to duplicate the study.

The methods section may contain several subdivisions. The experimental subjects and the criteria for including them in the study should be clearly explained. If patients are included in the study, a table of distinguishing characteristics may be helpful in summarizing demographic data. Such a table might include age, sex, diagnosis, mode of therapy, and so on. The sample selection procedure and the rationale for the study design must be clearly presented. This type of information is necessary if the results of the study are to be generalized to a broader population of similar patients.

An essential component of the methods section is a complete description of the equipment used to gather the data. The description should include the model number of the particular device and the manufacturer's name and address (city and state). For example, if airway pressure was measured, the text might read, "Airway pressure was measured with a differential pressure transducer (Validyne MP 45, 50 cm H_2O; Validyne Co., Northridge, CA)." The calibration procedure for each measuring device should

be described along with any pertinent validation procedures. Validation procedures might include comparison of a measurement device with a laboratory standard (e.g., using a spirometer to validate measurements made with a calibrated pneumotachograph) or testing the dynamic response of a device to assure accuracy at various frequencies (e.g., analysis of the frequency response of pressure and flow measuring systems).

The procedure used to gather the data should be described. This description might include an outline of the experimental protocol that was approved by the hospital's review board. State whether informed consent was obtained from the patient. A description of the experimental procedure should include the actual steps involved in gathering the data and the time elapsed during each phase of the experiment. Any problems or unforeseen events that occurred during the study should be mentioned. The information in this section should be detailed enough to guide other researchers who might wish to verify the results. The methods section will also help the reader to evaluate the quality of the data gathered during the study.

Finally, the report should include a brief description of how the data were analyzed. Provide a short discussion of the statistical procedures used and why they were appropriate for the experimental design. Unless the procedures were unusual, do not give the statistical equations used. However, many statistical procedures are based on certain assumptions about how the data were gathered (e.g., independence of data points used in a linear regression). Thus, enough information should be provided for the reader to evaluate the validity of any underlying assumptions and, hence, the adequacy of the analysis.

Results

This section of the paper presents the data gathered from the experiments. The order in which the information is given should correspond to the organization of the methods section. In that section, the reader was introduced to the step- by-step procedure used to study a particular problem. An expectation has been created in the reader's mind for the result of each step of the procedure. Therefore, the results should be presented in a logical progression from the beginning to end of the experiment. This progression helps to assure the reader of the thoroughness of the experimental technique.

The actual presentation of the data can take many forms. Use tables to summarize large amounts of raw data. Tables avoid the drudgery of long, redundant verbiage in the text of the paper. Each table should be constructed so that it’s meaning is clear without having to refer to the text. The idea is to summarize and guide the interpretation of large amounts of data and to reduce the time necessary to read the article. If the table appears to be too large, make use of figures or graphs. Specific examples of data summary techniques can be found in Chapters 10 – 13.

The information presented in the results section is usually in the form of "bare facts" with little or no explanation of its significance. Interpretation of the data is treated in the Discussion section. Of course, this is a general rule and may be suspended at times if it is felt that elaboration of some point would help the reader's flow of understanding. However, the overall train of thought created by a research paper should proceed from the presentation of the problem, to an explanation of the methods used to examine the problem, to the results of the experimental methods, and finally to an interpretation of the results.

From a philosophical standpoint, the results of a study do not "prove" anything. They simply add supportive evidence to the stated hypothesis. Thus, you must avoid any claims that a hypothesis was proven, confirmed, or verified in the results section. Remember that conclusions are described in the Discussion and Conclusions sections, not in the Results section.

The results of statistical analyses should be reported in the past tense. This is a subtle point. Consider, for example, the statement, "Airway pressures during high frequency ventilation *are* significantly lower than those during conventional ventilation." This implies that not only were the pressures analyzed during the study significantly different, but the author, by using the present tense, assumes that the results are unquestionably generalizable to all instances of high frequency ventilation. Using the past tense (changing "are" to "was") emphasizes that the results pertain to a limited sample under the conditions of one particular study. The responsibility for interpreting the generalizability of the results ultimately rests with the reader. The significance of a statistical test is usually reported in terms of a p value. Differences associated with p values less than 0.05 are considered significant by convention in medical studies.

Discussion

In the discussion, the author must show how the results answered the research question first described in the introduction. The results of statistical hypothesis tests must be translated into conclusions about the research hypotheses. This section allows the author to describe the implications and practical meaning of the study results. In addition, the discussion should describe the limitations of the study design, any problems encountered, and any recommendations for future studies. The process of interpreting the results concerns not only the data generated by the study but also relates that data to other studies and theoretical frameworks. The discussion is the appropriate place to include detailed reviews of other related research, including references, which would help to develop the reader's perspective and appreciation for the significance of the study.

Conclusion

Some journals require a separate conclusion statement at the end of the paper. The conclusions made should be thoroughly explained, including reasons for rejecting alternate interpretations. In addition, there should be a statement regarding the population to which the results can be generalized. Because the implications of a given study are usually speculative, it is appropriate to use words that are somewhat tentative in nature. For example, "The results of this study suggest that…" or "Because of the significant differences found, it may be possible to…". Such language emphasizes the fact that your interpretation is itself a hypothesis that may be tested by further research

Illustrations

The inclusion of a few well-composed illustrations can vastly improve the quality and readability of any manuscript. If drawings and photographs are to be included, it is usually best to consult a professional medical illustrator. However, with a few simple tools and a little practice, even a novice can achieve good results with graphs and simple charts. Simple line drawing can be done on a computer with a word processing program

Formal training is not necessary to create professional-looking graphs and charts. Simply look at examples of types of illustrations that appear in other manuscripts in various medical journals. Study the details and carefully, then copy that style using the current data. Having a manuscript published with original illustrations introduces a new dimension of satisfaction as well as economy.

SUBMISSION FOR PUBLICATION

First Steps

Read and follow the instructions to authors supplied by journal to which you will send your manuscript. The instructions will tell you things like how many copies to submit, how to format the references, how to submit electronically (if applicable).

Peer Review

After a manuscript has been submitted, the journal editor will look it over. If it passes this initial screening, it is coded and prepared for peer review. The editor may choose to return the manuscript with an explanation of why it is not acceptable. Sometimes the manuscript will simply not fit in with the typical subject matter of the journal, and the editor may suggest more appropriate journals.

Peer reviewers are carefully selected content experts; those with a thorough knowledge of the type of research being evaluated. Such experts have usually published extensively themselves. Ideally, the reviewers are not told the authors' names to help prevent bias.

Reviewing is hard work. Editors have responsibility to assemble knowledgeable reviewers and to protest when reviewers return an obviously biased or self-serving review. The reviewers have a responsibility to maintain their own integrity by looking beyond personal biases, to educate themselves about aspects of the study that may be unfamiliar to them, to weigh the quality of the paper with its potential value to the readership and to deal fairly with the author.

As an author, you can expect the reviewers to provide detailed and clearly written explanations of the strengths and weaknesses of your paper and support for any criticisms. Your paper will be judged (1) acceptable for publication as submitted (rare); (2) in need of revision before a decision can be made; (3) so flawed that adequate revision appears unlikely. An editor then gives consideration to the critiques of the two or three reviewers and informs the author.

Revision

Having a paper rejected is like being told your child is ugly and stupid. Most people react with negative emotions and give up. But if you can get past that phase, you have several options. First, examine whether the reviewers' comments are justified. Sometimes they have just misunderstood what you did. If the comments are justified, see what you can do to make the required changes. Keep two things in mind; (1) the time spent in revision is generally only a small fraction of the time already invested; (2) most manuscripts require revision and you are not being singled out. When the revision is complete, resubmit the manuscript with a cover letter listing each of the reviewers' comments and how you addressed them. An author always has the right to overrule a reviewer's objection, but he must adequately support his point of view.

Production

After a manuscript is accepted for publication, it is first copyedited. Copyediting is the stage at which editors make the manuscript consistent with their own journal's style. They may ask you to clarify some wording or supply missing information (e.g., incomplete reference data). When copyediting is complete, the manuscript is sent to be typeset. After it has been typeset, "galley proofs" or final page layouts are

printed and sent to the author for inspection and approval. Because of the costs involved, only minor changes can be made at this stage. When the author approves the galleys, he also grants the journal permission to publish the paper and assigns the copyright to the journal. The final stage of production involves careful proofreading (for typographical errors), pagination, transmission to the press, and the actual printing and binding.

MISTAKES TO AVOID

Here is a short list of some common reasons why manuscripts are rejected for publication

- The study did not describe the research question or it did not clearly explain the hypothesis.
- The paper did not actually test the research hypothesis
- The wrong measurement methods were used.
- The sample size was too small and the results were inconclusive.
- The study used the wrong design and did not adequately control for confounding factors.
- The statistical analysis was incorrect. Surprisingly, some papers even get published with the wrong analyses.
- The authors drew unjustified conclusions from their data.
- There is a significant conflict of interest (the authors might benefit form the publication of the paper and insufficient safeguards were see to be in place to avoid bias).
- The paper is so badly written that it is incomprehensible.

Chapter 16 on writing the case report goes into much more detail regarding common mistakes first-time authors make.

Here are some things to remember

- The single most important principle to remember is this: There must be a logical continuity throughout the different sections of the paper. The Title should hint at the hypothesis; the Introduction should state the hypothesis; The hypothesis should dictate the experimental procedures in the Methods; there should be data for all the methods in the Results section, and the Discussion should refer back to, and make conclusions, about the original hypothesis. Neophytes usually make the mistake of breaking this chain of continuity and the reader is left wondering what the paper is really about.
- Conduct the study without preconceived notions about the outcome. If you set out to "to prove that…" instead of "to find out whether…" it is hard to avoid personal bias.
- Always start your writing with the creation of an outline then follow the outline as you write.
- Supply all the detail necessary for other researchers to replicate your study. New authors sometimes keep all the necessary records but fail to include them in the paper.
- Repeat your measurements and report the number of repetitions. "Operator error" is a real concern and often can be detected by repeating measurements.
- Make sure your measuring devices are properly calibrated and document the calibrations procedures.

- Report only the data collected and draw conclusions based only on those data. Opinion can be included (sparingly) in the discussion but must be identified as such.
- Do a final proofreading of the manuscript after you have had a week or two (or even longer if possible) to forget about it. A fresh mind does the best proofreading.
- Submit a carefully prepared manuscript. Failing to consult the journal's instructions to authors and carelessness in preparation suggests that maybe the study was careless.

Writing for publication requires discipline, time and energy, but it completes the work of the investigator. It makes your work part of medical history. It is worth the effort!

QUESTIONS

True or False

1. Peer review means manuscripts submitted for possible publication in a medical journal are first reviewed by a panel of experts on the topic of the paper.
2. Authorship is based entirely on who collected the data.
3. You should always start writing your paper with an outline.
4. The basic format of a scientific paper is Title, Abstract, Introduction, Methods, Results, Discussion, in that order.
5. The single most important thing to remember about writing a research paper is that there must be a logical continuity throughout the Title, Introduction, Methods, Results, and Discussion sections.
6. Describing the calibration of measurement instruments is really not an important issue because readers assume your devices work properly.
7. Journals do not usually publish instructions for authors.

Chapter 15. The Abstract

As we learned in the last chapter, the abstract is a condensed version of a research paper that appears at the beginning of the publication. Many readers skim the abstract to see if they are interested enough to read the whole paper. Some readers don't have enough time to read anything more than abstracts. For these reasons, the abstract is an important element of a published paper. In fact, it is so important that sometimes it is published without the paper!

BACKGROUND

Many medical associations (including Respiratory Care) hold annual scientific conventions during which researchers give presentations of their work. The presentations are of two basic types: lectures and poster presentations. During the time leading up to the convention, the medical association will send out a "call for abstracts" requesting that researchers submit the results of their unpublished studies in abstract form. These abstracts are then subjected to peer review. Accepted abstracts are then published together in one issue of the medical association's professional journal. Authors of accepted abstracts are then invited to present their work at the convention in the form of either a lecture or a poster presentation. For the annual international respiratory care congress, all accepted abstracts are presented as posters. We will describe poster presentations in the next chapter. In this chapter, we will describe the type of abstract that is published in Respiratory Care journal, which is very similar to other medical journals.

Specifications

An abstract may report (1) an original study, (2) the evaluation of a method, device, or protocol, or (3) a case or case series. Topics may be aspects of adult cute care, continuing care/rehabilitation, perinatology/pediatrics, cardiopulmonary technology, or health care delivery. The abstract may have been presented previously at a local or regional (but not national) meeting and should not have been published previously in a national journal. The abstract will be the only evidence by which the reviewers can decide whether the author should be invited to present a poster at the annual meeting. Therefore, *the abstract must provide all important data, findings, and conclusions.* Give specific information. Do not write such general statements as "Results will be presented…" or "Significance will be discussed…".

Content Elements

Original study. Abstract *must* include (1) Background: statement of research problem, question or hypothesis; (2) Method: description of research design and conduct in sufficient detail to permit judgment of validity; (3) Results: statement of research findings with quantitative data and statistical analysis; (4) Conclusions: interpretation of the meaning of the results.

Method, device or protocol evaluation. Abstract *must* include (1) Background: identification of the method, device, or protocol and its intended function; (2) Method: description of the evaluation in sufficient detail to permit judgment of its objectivity and validity; (3) Results: findings of the evaluation; (4) Experience: summary of the author's practical experience or a lack of experience with the method, device, or protocol; (5) Conclusion: interpretation of the evaluation and experience. Cost comparisons should be included where possible and appropriate.

Case report. Abstract *must* report a case that is uncommon or of exceptional educational value and must include (1) Introduction: relevant basic information important to understanding the case; (2) Case Summary: patient data and response, details of interventions; (3) Discussion: content should reflect results of literature review. The author(s) should have been actively involved in the case and a case-managing physician must be a co-author or must approve the report.

Format

Abstracts are usually not typeset like papers. Rather, they are photocopied and printed in groups of four to the page. That means the abstracts must all conform to specific size specifications. Respiratory Care journal supplies authors with a form that they can use to type their abstract on. Alternatively, you can use the margin settings indicated on the form and type your abstract on a clean sheet of paper. For example, Respiratory Care requires the text box to have dimensions of 18.8 cm (7.4") by 13.9 cm (5.5"). The form is reproduced below.

A font like Helvetica or Times makes the clearest reproduction. Helvetica is easier to read but is slightly larger. Use Times if you need every bit of space. But don't use a text smaller than 10 points. Tables and illustrations must fit within the specified text area. No attachments are allowed. Standard abbreviations may be employed without explanation; new or infrequently used abbreviations should be spelled out on the first use. Any recurring phrase or expression may be abbreviated.

In some ways, writing this type of abstract is more difficult than writing a whole paper. You must think carefully about what you must include and what you would like to include if space permits. The best to write this type of abstract is to first read some published abstracts and study the ones that look neat and are easy to read and understand. Some good examples are included at the end of the chapter. Then create a rough draft of your own abstract on a clean sheet of paper. Don't worry about the size of the abstract but keep it to one page (using 12 point Helvetica font). Start, as always, with an outline (and one with more detail then just Introduction, Methods, Results, and Discussion).

After the first draft is finished, go back and consider how you might reorganize it to make it shorter. Start thinking about the final size. Rewrite the abstract so that it is about the right final size. Don't start new paragraphs for each section; that takes up too much space.

Next, start proofreading each sentence. Try to think of ways to reword sentences to make them shorter. Use shorter words and abbreviations where possible. Make your tables and illustrations as small as possible. This should get you to the final size of the required text box.

If you are still a sentence or two over the limit, switch to 10 point Helvetica. If you need just a little more, use 10 point Times. Make sure you bold the major headings (Background, Methods, etc.) so they stand out. Make sure your title is brief and correctly formatted (all bold, capital letters) so that it stands out too. Finally, make sure you submit the abstract following all the instructions to authors provided by the journal. And don't miss the deadline!

TEMPLATE

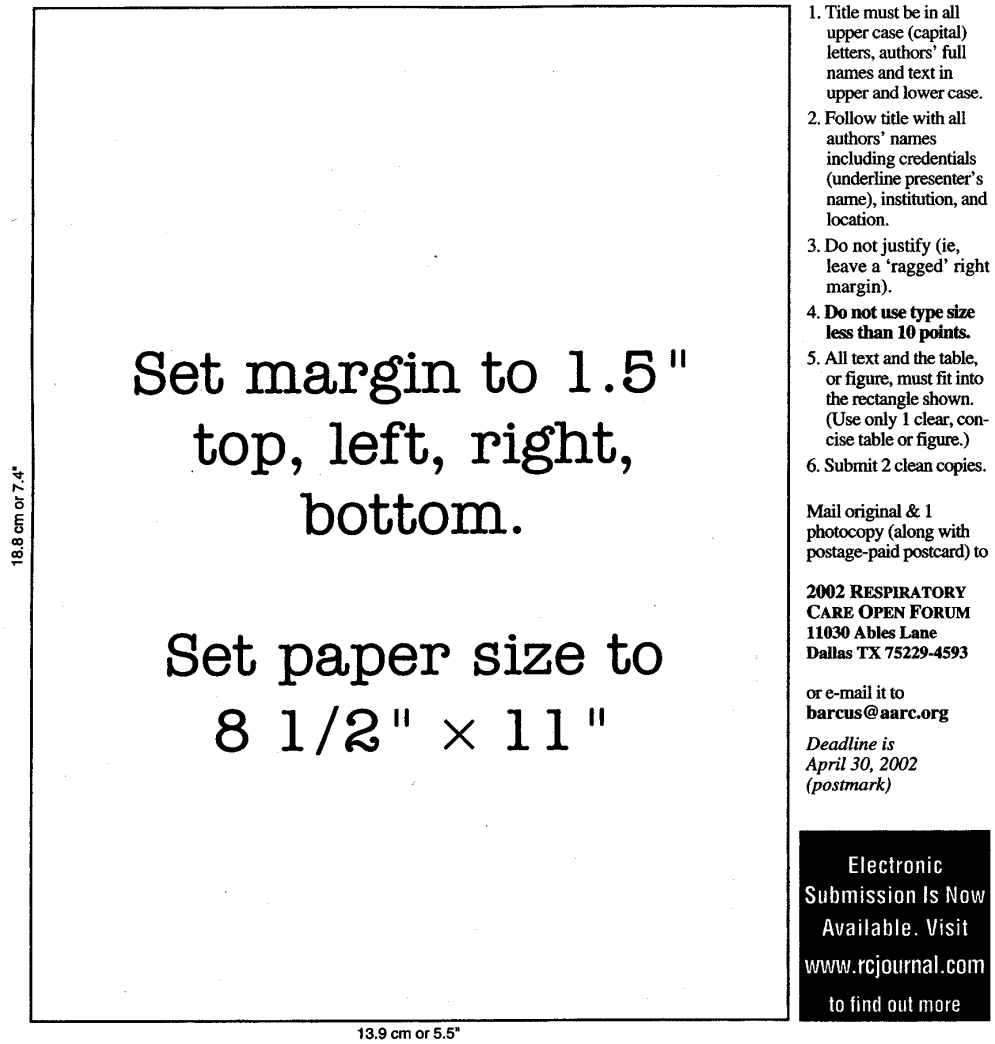
On the next page you will see a template illustrating the layout required for abstracts submitted to Respiratory Care Journal. You can find this in the back of any issue of the journal.

MODEL ABSTRACT

Following the template is an example of a well-written abstract that was submitted, reviewed, accepted and published in Respiratory Care. Study it carefully and note that it has all the elements required of a good abstract.

EXAMPLE TEMPLATE FOR SUBMITING AN ABSTRACT

RESPIRATORY CARE OPEN FORUM 2002 Abstract Form



1. Title must be in all upper case (capital) letters, authors' full names and text in upper and lower case.
2. Follow title with all authors' names including credentials (underline presenter's name), institution, and location.
3. Do not justify (ie, leave a 'ragged' right margin).
4. **Do not use type size less than 10 points.**
5. All text and the table, or figure, must fit into the rectangle shown. (Use only 1 clear, concise table or figure.)
6. Submit 2 clean copies.

Mail original & 1 photocopy (along with postage-paid postcard) to

2002 RESPIRATORY CARE OPEN FORUM
11030 Ables Lane
Dallas TX 75229-4593

or e-mail it to
barcus@aarc.org

Deadline is April 30, 2002 (postmark)

Electronic Submission Is Now Available. Visit www.rcjournal.com to find out more

Presenter	Name & Credentials Mailing Address Voice Phone & Fax
Corresponding Author *If Different* from Presenter	Name & Credentials Mailing Address Voice Phone & Fax

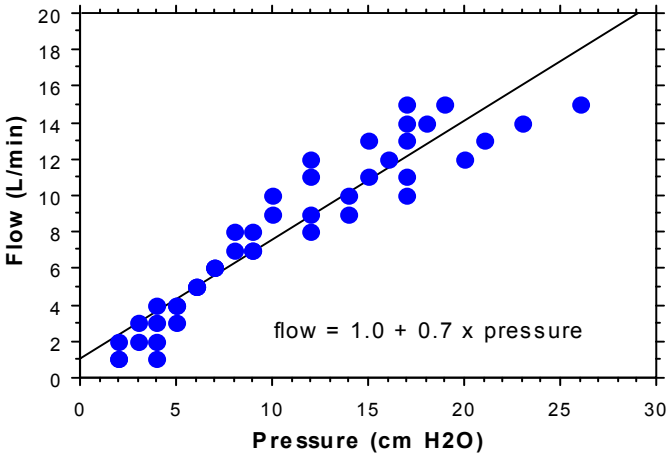
MODEL ABSTRACT #1

Respir Care 2001;46(10):1089

PRESSURE/FLOW CHARACTERISTICS OF THE EZPAP POSITIVE AIRWAY PRESSURE THERAPY SYSTEM

Snyder, RJ RRT, Slaughter, SL RRT, Chatburn, R RRT, FAARC
University Hospitals of Cleveland, Cleveland, Ohio

BACKGROUND: The EzPAP positive airway pressure therapy system is indicated "for lung expansion therapy and the treatment and prevention of atelectasis". The device generates positive airway pressure using flow input from a standard air or oxygen flowmeter and may be used with a nebulizer. The purpose of this study was to evaluate the pressure/flow characteristics of the EzPAP with and without a nebulizer. **METHODS**: The gas inlet port of the EzPAP was connected to an air flowmeter. The pressure monitoring port was attached to a pressure gauge calibrated with a water manometer. We measured the inspiratory and expiratory pressures with 3 normal volunteers inhaling and exhaling normally. Input flows were adjusted in 1 L/min increments from 0 to 15 L/min. The experiment was repeated with a nebulizer attached. The nebulizer was powered by a flowmeter set to 6 and 8 L/min. Each experimental condition was evaluated 3 times. Data were analyzed with linear regression. **RESULTS**: The addition of a nebulizer had little effect on the airway pressures generated. The data below show airway pressure as a function of the flow set on the flowmeter for all experiments combined. The coefficient of determination (r^2) was 0.88. The system maintained airway pressure throughout the respiratory cycle with the inspiratory pressure averaging 45% of expiratory. The chart illustrates expiratory data.



EXPERIENCE: The EzPAP was very easy to setup and adjust. Breathing on the device was similar to a PEP device except that inspiratory pressure was elevated. The volunteers commented that they felt a noticeable increase in FRC. With the use of higher pressures/flows, subjects felt they needed to actively exhale. **CONCLUSIONS:** The EzPAP delivers reliable and controllable pressures within a clinically useful range for lung expansion therapy. An easily remembered equation allows prediction of the flow necessary to achieve a desired pressure.

MODEL ABSTRACT #2

Respir Care 2001;46(10):1127

CRITICAL PATH ANALYSIS OF A CYSTIC FIBROSIS CARE PATH IDENTIFIES PROCESS IMPROVEMENT OPPORTUNITIES

Teresa A. Volsko BS, RRT; Scott Grey BS; Robert Chatburn BS, RRT;
Sally Lambert RN, PhD; Michael W. Konstan, MD
Rainbow Babies and Children's Hospital, Case Western Reserve University, Cleveland, Ohio.

Background: Network diagrams and critical path analysis are used as resources to manage complex projects. These project management tools show process interdependencies, identify rate limiting steps and streamlining opportunities, and estimate a time line for project completion. We used these tools to construct a care path to improve the care of patients (? 18 yrs.) with cystic fibrosis hospitalized for an acute pulmonary exacerbation. We hypothesize that the care path length of stay (LOS) predicted by the critical path analysis would match actual patient data and the analysis of key patient care processes in the critical path could be used to identify performance improvement areas. **Methods:** Four key patient care processes were identified. Two were streamlining opportunities (physician order entry and initial antibiotic administration time). The remaining were rate-limiting steps (obtain patient sputum sample and report culture and sensitivity results). Standard times were estimated from physician and staff expert opinion (order entry time, time to obtain sputum specimen). Current hospital pharmacy and laboratory time standards were used as standards for initial IV antibiotic administration time and culture and sensitivity reporting time. A network diagram was constructed to map patient care from the time of admission to the completion of the key care processes. The duration of the critical path, in days, was the projected LOS. A care map was constructed to standardize the care path's laboratory tests, and diagnostic and therapeutic modalities over that projected LOS. The care path was implemented and data prospectively gathered from January 1, 2000 to December 31, 2000. Key process and outcome (LOS) data were collected into a database and analyzed using one-sample *t*-tests, where the specified test value was set to the projected standard times for each key process and the projected LOS from the care path model. Statistical significance was set at $p < 0.05$. **Results:** Fifty-eight patients meeting the care path criteria were treated over the 12-month period. The table below compares the projected and actual measured mean times for each key process and the LOS for the care path:

Key Care Processes in the Critical Path	Projected Standard	Measured Value mean (± SD)	p Value
Physician orders entered and activated	1 hr.	0.95 hrs. (0.77)	0.59
Initial IV antibiotics administered	2 hrs.	3.12 hrs. (1.36)	< 0.001
Sputum specimen obtained	2 days	1.27 days (2.4)	0.03
Culture and sensitivity results	5 days	7.2 days (3.73)	< 0.001
Length of stay	9 days	10.03 days (6.3)	0.22

The LOS and three key processes predicted by the critical path matched or were less than actual patient care data. However, two key steps of the critical path (initial IV antibiotic administration and culture and sensitivity results) exceeded the projected standards, and present opportunities for additional improvement. **Conclusions:** A well designed care path, using network diagrams and critical path analysis, can predict actual patient care outcomes and provides a mechanism to identify, measure and improve key patient care processes.

WHAT NOT TO DO (ANALYSIS OF REJECTED ABSTRACTS)

The following pages show copies of abstracts I have seen as a reviewer and rejected for a variety of reasons. They are perfect examples of what not to do. Read each one through as if you were a reviewer and try to spot the problems. In the Appendix, you will find checklists that reviewers for Respiratory Care Journal use to appraise papers and manuscripts. Use the appropriate checklist (original study, device evaluation, or case study) to judge the quality of the abstracts in this section. After each abstract, we will discuss some of the things I saw.

Note: pay no attention to the size of the abstracts, requirements have changed over the years and none of these was rejected because of being the wrong size. Note also that the author and institution names have been removed.

USING CQI TO IMPLEMENT THERAPIST DRIVEN PROTOCOLS

One of our CQI monitors in the Children's Hospital reviewed the oxygen saturation (SpO2) of patients receiving O2. This monitor revealed that patients consistently received prolonged administration of excessive O2 with inefficient weaning taking place. The criteria for "excessive O2" was a consistent SpO2 > 96%. The CQI Committee felt that this situation was due to a lack of structure in our weaning process and proposed the development and implementation of a therapist driven O2 weaning protocol. A literature search on weaning pediatric patients from oxygen was conducted. Also, attending physicians were surveyed about how they would like to see their patients weaned. Twelve of 22 physicians responded indicating they would like to see their patients checked/weaned at least every 3 hours for saturations >95% in approximately 5% decrements. Using this information an O2 weaning protocol for the general care areas was written. The goals of the protocol were to allow for a more efficient use of O2, to improve patient care and to reduce patient cost. The weaning protocol was approved by the Respiratory Therapy Medical Director and by the Medical Executive Committee. The protocol was implemented in November, 1993. Following a random selection of patients with diagnoses which precluded the necessity of chronic O2 administration, such as BPD and Cystic Fibrosis, 17 patients were selected for chart review from 1991, prior to the protocol, and early 1994, post protocol. Diagnoses reviewed were Asthma, Croup, Bronchiolitis, Pneumonia and Post-op patients. The average period on O2 for these patients in 1991 was 2.6 days. In 1994, the average period was 1.6 days. Although not stastically significant (p > .05), these preliminary findings indicate the protocol may be effective decreasing O2 use in these patients. The CQI process, when appropriately used, can result in potential cost savings and improved patient care.

Abstract #1

What is the very first thing you notice about this abstract? It is not in the standard format of Background, Methods, Results, Conclusion. The first impression you get is that the author(s) were careless in preparation. They were probably inexperienced and without a good mentor. Let's ask some basic questions:

Is the background information adequate? In fact it reads like its all background information. Just a narrative and not a study at all. We can't even tell if this is an original study or a device/protocol evaluation.

Is there a hypothesis or explanation of device? We can't really tell what the study question is and there certainly is no hypothesis statement. It just rambles on without any real focus.

Is there enough detail in the methods? Well, there was a literature search but we don't know why. There was some kind of survey but no detail of questions asked. Based on the survey a protocol was written and implemented. Patients were randomly selected but selection criteria are not clear and we don't know why they were reviewed until we see results on oxygen usage (ie, no prior hypothesis). Methods and results are scrambled together. A *p* value is given but we don't know what statistical test was performed.

Are results complete? We have to hunt for results guess at their validity.

Are conclusions appropriate? The only outcome variable seems to be period of oxygen use and there was no difference before an after the protocol. There are no data on cost or quality of patient care, so no conclusions can be made on cost or quality.

Verdict: This abstract is a disaster!

COMPARISON OF WHOLE BODY PLETHYSMOGRAPHY (WBP) WITH IN-LINE PNEUMOTACHOGRAPHY (ILP) FOR NEONATAL PULMONARY FUNCTION.

ILP, which has long been the standard device for air flow measurements in newborn pulmonary function, is known to be affected by barometric pressure, FiO_2, temperature, airway pressure and humidity.

Two ILP volume monitoring devices (Bear NVM1 and Bicore 100) and a WBP device (Vitaltrends VT1000) were compared to understand the affects of changing gas conditions under clinical settings. Tidal volumes (V_T), Inspiratory time (TI) and resistance (R) were compared.

Without BTPS and FiO2 correction, the effect on volume measurement was found to be large for pneumotachography ($VT_{VT1000}=1.24*VT_{Bicore}+.28$, $R^2=.999$, $VT_{VT1000}=1.18*VT_{NVM}+.31$, $R^2=.993$ at FiO2=100). When corrected for BTPS and FiO2 the Bear NVM1 and the VT1000 show the same volumes ($VT_{VT1000}=1.09*VT_{NVM}+.29$, $R^2=.993$ at FiO2=100, $VT_{VT1000}=.93*VT_{NVM}$, $R^2=1$ at FiO2=21) The VT1000 calibrates itself for FiO_2 and BTPS. The Bear NVM1 provides a lookup table for BTPS corrections. When not taking into account the difference in mean level the interclass correlation coefficient (ICC) was found to be significant in all cases (ICC=.98, .97 and .98 for VT1000 vs NVM1, VT1000 vs Bicore and Bear vs Bicore respectively). Pulmonary mechanics measurements were similar between the Bicore and the VT1000 (ICC=.99 for TI, ICC=.9 for R). Although resistance trended similarly there was a large offset.

When uncorrected as is typically the case in clinical settings, the ILP method can over or underestimate minute volume by as much as 24% in our study. We conclude that WBP measurement is accurate and convenient for monitoring critically ill neonates, and provides a more accurate bedside readout of minute volumes under conditions of varying FiO2, RH, temperature and atmospheric pressure.

Abstract #2

Again, the first thing you notice about this abstract is that it is not in the standard format of Background, Methods, Results, Conclusion. At least these sections are not clearly identifiable.

Is the background information adequate? Maybe too brief but it is there.

Is there a hypothesis or explanation of device? There is no definite hypothesis statement but we can piece together the idea that the authors wanted to evaluate the effects of gas condition on the accuracy of tidal volume, inspiratory time, and resistance measurements. You wonder why those particular measurements were chosen.

Is there enough detail in the methods?

What methods? The authors jump right into results. Are the measurements made on patients or a lung simulator? What were the gas conditions (eg, temperature, barometric pressure, FiO_2)? A correlation coefficient is the wrong statistic to compare agreement among measurements (see chapter on basic statistical procedures). Regression may be alright in this case but we only know the method was used if we are familiar with the format of the equations presented as results.

Are results complete? Plenty of numbers but we are not sure how they were obtained.

Are conclusions appropriate? The authors assume the VT1000 is the standard (again we have to guess this from the format of the regression equations). Why should we believe that assumption? They conclude that WBP is accurate but it was not compared with any type of "gold standard" to make that conclusion. Indeed, it was only compared with devices expected to be inaccurate due to lack of correction factors.

Verdict: This abstract was a little better than the last but still deplorable.

EFFECT OF PEEP IN CONGENITAL DIAPHRAGMATIC HERNIA

Introduction: The high risk congenital diaphragmatic hernia (CDH) infant continues to be a challenge to ventilate. Little has been published on the use of PEEP in these infants post-ECMO. We evaluated the effects of PEEP during trials off of ECMO on lung compliance (Cdyn), physiologic deadspace (VD/VT), $PaCO_2$ and PaO_2 in CDH infants. **Methods:** Patients were sedated, paralyzed, and ventilated in the pressure control mode on the Servo 900C. Standard ventilator settings were a PIP/PEEP of 30/5 cm H_2O and a rate of 30 breaths/min. Cdyn, VD/VT ratios, and arterial blood gases were obtained during routine separations from the ECMO circuit. PEEP levels were lowered to 2 cmH_2O while maintaining the same peak inspiratory pressure. Measurements were then repeated. **Results:** Seventeen CDH infants who required ECMO were evaluated. There was a significant improvement in Cdyn ($p=.006$), VD/VT ($p=.011$), and Vt/kg ($p<.001$) when decreasing the PEEP from 5 cmH_2O to 2 cmH_2O. The $PaCO_2$ also significantly improved on a lower PEEP ($p=.02$), however, there was no significant difference in the PaO_2. **Discussion:** We conclude that both VD/VT and Cdyn in the CDH infant significantly improves on low levels of PEEP. Concurrently, the lowering of the mean airway pressure does not adversely affect the PaO_2. This suggests that PEEP levels greater than 2 cmH_2O worsens physiologic deadspace and compliance during trials off of ECMO potentially altering the clinical assessment. Future studies will be directed to determine whether it is the anatomic and/or alveolar deadspace that increases with higher PEEP levels.

Abstract #3

At last we find an abstract with the correct format; Background, Methods, Results, Conclusion.

Is the background information adequate? Yes, short, relevant, and to the point.

Is there a hypothesis or explanation of device? Study purpose is stated but it would have been clearer if a specific hypothesis was given.

Is there enough detail in the methods?

No description of study entry criteria for patients (some info is in the Results section). How many times were measurements repeated? How were deadspace and dynamic compliance assessed? The mode of ventilation is not clear. No description of statistical procedures.

Are results complete? No actual data are given. We would like to have seen summary data at least like mean and standard deviation values for measured variables.

Are conclusions appropriate? Assuming the statistical procedures were done correctly, the conclusions are justified.

Verdict: Without more detail in the methods and some actual data, we cannot judge the value of this study. The authors ask us to take too much on faith. Close, but no cigar!

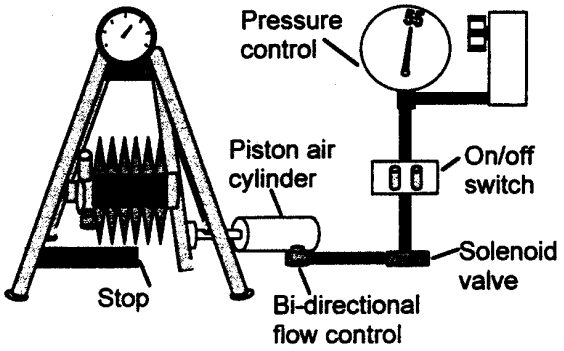
Abstract #4

This last example is very interesting. Read it over carefully:

SIMULATION OF CLOSED CHEST COMPRESSION ON MECHANICAL TEST LUNG

BACKGROUND: Closed-chest compression during cardiopulmonary resuscitation (CPR) is an important lifesaving procedure.[1] Breathing assistance devices, such as manual bag-valve and automatic resuscitators, are pressure sensitive, and their function is affected by chest compression. As breathing assistance devices evolve, the need exists to test these devices, techniques, and equipment to ensure their safety and efficacy. Such a system is described below.

EQUIPMENT: The lung environment was simulated by a commercially available lung simulator (SMS, England MS0015001). An external compression simulator was attached to its structure as well as a pressure tap for data acquisition. The simulator was equipped with controls to regulate cycle rate, speed, force, and distance. The lung compression device was comprised of an air cylinder with bi-directional flow adjustments to control compression speed, a pressure adjustable air supply to control maximum force, and control circuits and a solenoid valve for actuation. The cylinder mount directed the cylinder rod against the mechanical test lung and a stop was mounted to the frame to simulate the maximum depth of an external compression. A fine adjustment scale on the speed and pressure control was essential to ensure repeatability.



METHODS: A breathing device, such as a manual bag-valve or automatic resuscitator, was used to simulate inhalation and exhalation functions. The chest compression simulator was cycled to simulate CPR. This was helpful in evaluating synchronous and non-synchronous compressions with breathing cycles at various compression rates.

RESULTS: The compression rate was set at 80 times per minute with a ratio of 5 compressions per breath. The proximal airway pressure waveform was recorded utilizing a computer data acquisition system. The pressure waveform represented external chest compressions during resuscitation.

CONCLUSION: Because this lung compression system is so versatile, it can be used to evaluate many different types of respiratory devices. In this case, closed chest compressions were simulated under controlled conditions. This system can be effectively used to test the safety and efficacy of other breathing assistance devices.

Supported by VORTRAN Medical Technology 1, Inc.

1. National Conference on Cardiopulmonary Resuscitation and Emergency Cardiac Care. Standards and guidelines for cardiopulmonary resuscitation (CPR) and Emergency Cardiac Care (ECC). JAMA 1986;255(21):2905-2984.

So what do you think? It's beautiful, isn't it? Great layout, easy to read, good introductory sentence with the advanced feature of a reference to other published work. There is even a nice line drawing. There is only one problem; this is not a device evaluation. It is just a narrative description of a device. A nice little story, but not even close to being a scientific study.

SUMMARY

By now, you should be familiar with what an abstract is and how to create one the proper way. The mistakes that novices make have been illustrated and you have seen what good abstracts look like. Abstracts are important because for many people, they are the first (sometimes only) involvement in research publication. And of course, once the abstract is written and published, most of the work for a full-length paper has been done.

Keep in mind that even though most journals subject abstracts to peer review, they still do not have the same credibility as papers. This lack of credibility is because the methods are necessarily abbreviated and readers cannot properly judge the study's validity. Many bad quality abstracts slip by the reviewers due to the sheer numbers submitted and published. Always interpret any abstracts you read with a healthy skepticism.

QUESTIONS

True or False

1. You may publish only an abstract of a study and not write a whole paper.
2. Usual topics for an abstract are (1) an original study, (2) a device or method evaluation, and (3) a case study.
3. The abstract should give a brief overview of the study and not include any data.
4. A published abstract is just as credible as a published paper.
5. Abstracts are never peer-reviewed like papers are.

Chapter 16. The Case Report*

Inclusion of a chapter on writing case reports might at first seem out of place in a text on clinical research. Properly selected and prepared, however, the report of a single case cannot only qualify as legitimate clinical research but also present an opportunity for health care practitioners to participate in the research and publication process when circumstances render other forms of research publication unfeasible. One need not have research funding, a laboratory, or a "sabbatical" from one's patient-care-based hospital position in order to prepare and publish an excellent case report. In addition, experience with a case report can help to create, in someone without previous research experience, the kind of critical frame of mind that can be the crucial ingredient missing in unsuccessful attempts at formal research endeavors.

Virtually everyone participating in the assessment and care of patients with respiratory diseases encounters findings or outcomes, or solves clinical problems, in ways that could be "reportable" if they thought about them in the right way and then were able to follow through with "writing them up" correctly. The purpose of this chapter is to stimulate health care practitioners to think about the possibilities of investigating, learning from, and sharing with others in their field their experiences with patients, and to provide guidelines for turning those experiences into publications in the scientific literature. Case reports seldom make medical history. In general they are less significant scientifically than original investigations. On the other hand, there are some kinds of scientific information that cannot conveniently be communicated in any other way. The case report can be an important link in the larger chain of scientific progress and communication. Certainly, a case report must have the same sophistication, honesty, authoritativeness, and professionalism as a prospective study or comprehensive literature review. Thus, although it is one of the lesser players on the scientific stage, the case report requires no less care and must meet the same standards as its more celebrated colleagues in print.

Most case reports consist of a description of an individual illness, course, or treatment, followed by a discussion of the unique or particularly instructive features it demonstrates. In the majority of instances such a report is a few pages long at most and contains perhaps one figure or table and half a dozen references. There are, however, other formats that may occasionally be acceptable. Two cases may be reported, or even four or six, using the same basic outline, with discussion of the primary feature being reported coming after the last case description. The difference between such a multi-case report and a clinical series lies mainly in the former is anecdotal, and thus less scientifically rigorous, nature. Another format is the "case report and review of the literature," which consists of an extensive review of all the important published experience with a subject that is introduced by an illustrative individual case. In comparison with the usual, shorter type of case report, these have fewer opportunities for publication, since most medical journals today do not have the space to publish them.

A last variant of the basic case report is the case study, which is often a lengthy description of detailed laboratory investigation and represents highly original research, even though the research may be directed to elucidation of the disorders of a single patient. Detailed investigation of a single case is no substitute for a carefully designed clinical trial; instead, the instructive value of an individual case can

* This chapter was written by David Pierson, MD and originally appeared in a book called *Fundamental of Respiratory Care Research* by RL Chatburn and KC Craig, now out of print.

sometimes be increased by careful modification of the circumstances under which the observation in question was made.

This chapter first discusses the qualifications for authorship of case reports, and then summarizes the types of cases or observations that might be appropriate subjects for a published case report. Next comes a step-by-step discussion of the process of preparing and writing such a report, and a summary of its usual components. Finally, a dozen mistakes are described that authors of case reports, especially first-time authors, most frequently make. Ways are suggested to avoid these mistakes.

WHO SHOULD WRITE IT?

The case report can be an ideal "first paper ". Most reports of original investigation require expertise in several areas of scientific endeavor not ordinarily within the experience of the health care practitioner. But the case report can be written by anyone in the field, so long as he or she is prepared to work hard at it and obtain the right kinds of help. Writing a case report requires that the author thoroughly understand the case being reported and the disease or entity it is intended to illustrate; these requirements in turn imply access to reference materials and other facilities necessary to substantiate them.

A report of a new clinical finding or disease manifestation may well require expertise beyond that of most respiratory therapists, and simply reading about the conditions involved may not be sufficient for the task of writing the paper. In most circumstances, help will be required from someone with clinical experience in the relevant area, and the source of such assistance may not be immediately apparent. The medical director of health care services, another attending physician, or an instructor would be logical persons with whom to start, as would anyone else with personal experience with the condition in question.

Of course, a paper written for publication and the eyes of one's peers must be written in clear, correct English, and expertise in English composition is not necessarily a requirement for clinical work. Thus, the writing aspect of the case report may require help from someone more experienced, especially if English is not one's strong suit. This mundane aspect of the publication process is nonetheless vitally important, and medical journals often reject poorly written manuscripts irrespective of their scientific interest or accuracy.

There is no reason that one person cannot recognize, research, prepare, and submit a case report for publication. To do so requires expertise in the clinical practice, in literature review, and in scientific writing. Thus, most case reports, and especially those of first-time authors, have several contributors, whose skills in different areas make the end product better than it would be with a single author. In most clinical reports, quite a number of individuals have been involved during various phases of the case. These may include the attending physician, consultants, nurses, respiratory therapists, laboratory personnel, statisticians, librarians, and secretaries. Who should and should not be listed as an author of the paper? This will vary depending on the complexity of the case, but as a rule most case reports should not require more than three or four authors. If more names than this appear on the title page, it is likely that one or two individuals really did the lion's share of the work; listing the others in effect detracts from the credit they deserve. Whoever had the original idea, and actually writes the paper, should be first on the author list. The first listed author was usually responsible for the project and did more work than anyone else. Other individuals who make major contributions to the conception, background work, and writing should be co-authors. But do not include persons whose involvement was limited to taking care of the patient, running laboratory tests (unless these are experimental or special research procedures), interpreting x-rays, and so forth. Some department supervisors or medical directors insist

that their names be on any paper to come out of their departments. This practice is unprofessional unless these individuals also make major contributions. Deciding early on who is and is not to be an author will save the project from later misunderstandings and unpleasantness.

ATTRIBUTES OF A REPORTABLE CASE

Some editors looked down on Case reports as being less important contributions than papers describing prospective, controlled clinical trials and other kinds of research. In this attitude they are correct. Case reports seldom make large impacts on medical progress. Several medical journals, in fact, no longer publish case reports at all. However, most editors recognize the distinct value of an appropriately prepared and targeted case report, and most clinical journals publish at least some.

What cases are appropriate varies with the journal and with its editorial policy. One prominent editor has stated that there are three kinds of single-case reports that "still occasionally merit publication":

- the unique (or nearly unique) case, unreported previously,
- the unexpected association of two or more known conditions or diseases in a single patient,
- the unexpected outcome or event that suggests response to therapy or an adverse drug effect.

In a broader sense, cases may be worthy of publication if their description and discussion presents either something new or something already known but in a particularly instructive example. To be something new does not necessarily require uniqueness in all the world's literature. Rather, such “news” should be previously unknown or readily available within the health care setting. Also, a particularly instructive example may be defined differently in different situations.

The following is a discussion of six settings in which publication of a case report may be justified (Table 16-1)

Table 16-1. Possible reportable cases

- A new disease or condition
- A previously unreported clinical feature or complication of a known disease or condition
- A particularly instructive case of a previously known disease of condition
- A new diagnostic test or other assessment as illustrated by an individual case
- A new treatment modality
- A new outcome or complication of treatment

A New Disease or Condition

In recent history, the acquired immune deficiency syndrome (AIDS) has reminded everyone connected with health care that medicine is not static. New diseases and disease-causing agents continue to appear. Still, a truly unique syndrome or entirely new disease is pretty rare. However, "new" need not mean "previously unreported anywhere," but rather a condition pertinent to the field that has not previously been described in the journals most likely to be read by one's peers.

A Previously Unreported Feature or Complication

Unreported features might include situations documented previously but not in the usual literature of the health care practitioner. But those categories imply a burden of proof on the author to be certain that the disease or situation described has not in fact been similarly reported in the past. Thus, you must make as reasonably thorough a search of the literature as possible. Even when no prior description has been located, it is still unwise to make statements like "this is the first reported case," and few editors will allow such proclamations to be printed.

A Particularly Instructive Example of a Known Condition

A well-studied or unusually well-documented example of a known or even common condition can be a valuable teaching tool. Most medical journals do not publish formal case reports of such cases. Many journals have regular features such as "Case Records of the Massachusetts General Hospital" in *New England Journal of Medicin.* This feature is a weekly clinico-pathological conference that serves the same purpose. However, *Respiratory Care* and certain other journals may accept well-documented examples of known diseases or outcomes as case reports if they are authoritative and do not duplicate recently covered material.

A Case Illustrating a New Diagnostic Test or Monitoring Technique

Case reports in this category are exceptional. However, some methods of patient assessment cannot be initially evaluated with more rigorous trials involving large numbers of patients. A case report or small series of individual, anecdotal cases can be a reasonable alternative. For example, suppose persistent bronchopleural air leak during mechanical ventilation is an unusual occurrence at your institution. You have devised and documented a new, accurate way of quantitating the leak. A case report may constitute the only means readily available to disseminate information about it to your peers.

A New Treatment Modality

A management technique applied to an infrequently seen condition may have to be reported by means of a single case. The practitioner either sees too few of them or does not have the resources to perform a formal trial in a larger number of cases. This circumstance has admittedly been abused. The health care literature contains too many single case reports of new treatments that appeared to work but have never subsequently been validated with appropriately designed investigations.

A New Outcome of Treatment

An unusual result or complication of treatment can form the basis of a case report, particularly if the events reported constitute a dramatic departure from the usual or expected. If a patient survives a previously uniformly fatal condition, or dies from something that is not supposed to be serious, this should be documented. Documentation of adverse outcomes and complications is an important step in the eventual establishment of optimal patient care.

STEPS IN PREPARING A CASE REPORT

The principles of research design are applicable in many respects to the preparation of a single case report. In general, however, the process is less complex and some phases may be unnecessary. Preparing a case report is a project of much less complexity and magnitude than most formal research projects. The difference is mainly because only a single patient or event is being described and the topic is

intentionally a narrower one. Table 16-2 lists the stages through which most case reports pass, from initial conception to ultimate publication.

Table 16-2. Steps in writing a case report.

- Identification of an appropriate case
- Review of the pertinent literature
- Consultation and discussion
- Planning the paper; assignment of roles and authorship
- Further investigation of the patient, institution of new therapy, or other original research
- Preparation of the first draft
- Preparation of tables and illustrations
- Consultation and discussion
- Revision of the manuscript
- Preparation and submission of final draft

Identification of an Appropriate Case

Health care is a dynamic, rapidly evolving field that deals with diseases and treatments that were unheard of a generation or two ago. In this setting, the clinician inevitably encounters clinical presentations, complications, or outcomes that he or she has not previously seen. It is safe to predict that during a career in clinical health care, each of us will encounter cases or events that could be "reportable". We thus have the opportunity to contribute to the state of our knowledge in our field. In the great majority of instances, however, the unique or innovative will be overlooked. The cases reported in *Respiratory Care, Chest,* and *Anesthesiology* are not the only ones that happen; they are the ones that are recognized and pursued in the spirit of scientific investigation. Thus, the vital ingredient enabling you to make the first step toward publishing a case report is a continual watchfulness for the interesting, the stimulating, and the different from what one usually sees. In most instances, cases so identified turn out to be instructive but already known to others in the profession. To find that one in five or one in twenty that could and should be pursued and published, however, it is important to establish an attitude of intellectual inquiry in one's work. You should be thinking constantly of how one might learn more from everyday experience.

Review of the Pertinent Literature

Once a potentially new situation is encountered, you must learn more about the subject area in which it falls. Even "experts" who have performed research on the subject cannot say for certain that a particular circumstance hasn't been reported before. A literature search is always necessary. Start with textbooks or journals so that undue effort is not expended on expensive computerized searches. But if no previous reference to a similar case is found, a thorough literature search must be done.

Consultation and Discussion

Before too much energy is put into a project, you should consult advisors whose experience is more extensive than your own. A therapist or technician might well discuss the case with the medical director of the department, or with a specialist in the area involved, or with an instructor in a local training program. These individuals will not necessarily become co-authors should the case proceed to that stage, but their perspective can help to direct the author's efforts and to avoid unnecessary labor. If no such individual is readily available locally, consider writing a letter to the editor of the journal to which you want to submit. You could also contact one of the individuals listed as a consultant or editorial board member on the journal's title page.

The physician caring for the subject of the case report must be informed that his or her patient's case is being considered for publication. In some cases, the patient's physician should be a co-author. But this is not usually the case unless that physician initiated or participated in the study that is being described in the report. Instances have occurred in which two or more separate groups of individuals have unknowingly prepared manuscripts reporting the same case. This can happen when the case has attracted a lot of attention or involved a large number of consultants.

Planning the Paper and Assignment of Roles and Authorship

Once the prospective author has become something of an expert on the subject of the report, the "nuts and bolts" of the paper should be planned. The content of each of the structural elements of the case report (described in subsequent sections of the chapter) should be outlined. If photographs, artist's drawings, or statistical analyses will be required, they must be no less professional than the rest of the paper, and appropriate facilities must be located.

Most case reports have two to four authors, each of whom plays a different role in assembling the components of the final product. Who is and is not to be listed as an author should be agreed upon as early in the process as possible in order to avoid later conflict and antagonism. To be an author, each person must play a vital role, not necessarily in the care of the patient but in making the case report happen. Authors are those who are instrumental in library research, writing, editing and preparation of tables or illustrations. It is also helpful if a rough schedule can be drawn up at this time. The participants should agree to meet regularly until the paper is finished.

Further Investigation of the Case

Case studies are based primarily on the value of carefully designed scientific investigation carried out on a particular patient once the uniqueness or other value of that patient's illness has been recognized. In such instances, all the steps described thus far consist of preparation for additional data collection based on what has been learned to that point. Most case reports do not contain this step of further investigation. But in those that do, it must be thoroughly thought out and designed before any more data are collected. Simply ordering all the tests one can think of on a patient with an unusual condition will waste resources, subject the patient to needless risk, and be unlikely to produce the new knowledge required for a worthwhile case study. The collaboration of an expert with experience in investigating similar cases is vital to the success of this type of case report. Data collected prospectively in the in-depth study of a single patient must be handled like data collected in other types of clinical research, as described elsewhere in this book.

Preparation of the First Draft

Writing the first draft is the most difficult step for first-time authors and for many experienced scientific writers as well. You can get “unstuck” by simply striving to get *something* down on paper, to get the process rolling and to provide something on which to build. One effective approach is to first all your data and notes, assembled during background reading and meetings among the paper's authors. Then make a one-page outline of the paper, including major components and points to be covered. Next, sit down and work your way through the pile of notes, making a crude narrative that follows the outline. At this stage, you should pay no attention to grammar or punctuation, but simply try to get it all down. This sloppy, too long, usually confusing document constitutes a first draft. The draft serves as the real foundation on which the final manuscript will be constructed.

There are many other good ways to write a paper. Someone who has never written a comprehensive term paper or formal essay in high school or college might well enlist the help of an individual with more experience, even if they were not involved in the case being reported. Several excellent books on basic scientific writing can also be helpful to first-time and experienced authors alike.

Preparation of Tables and Illustrations

Most case reports require means other than tests to convey important aspects of the case and to clarify their instructional value. This is not to say that a table must be assembled, or x-rays photographed, just to "flesh out" the report. Tables and illustrations can be difficult to prepare correctly, and including them is more expensive for the journal so each one must communicate information vital to the case that cannot be conveyed more effectively in another way. Tables and figures are not peripheral details to be taken care of as the final version is readied, but primary, central features of the paper. Along with the abstract, they will be the first parts of the paper to be looked at by reviewers, the editor, and the eventual reader.

Consultation and Discussion

After a first, complete, working draft has been put together, it is a good idea to have an experienced individual not connected with the project examine the entire draft. This may not be required if one of the authors has extensive experience both with the subject matter and with manuscript preparation, but can help to add perspective and guide revision and thus make the manuscript better. Having the paper "pre-reviewed" locally, before submission to a journal and while it can still be revised, can improve its chances for acceptance in the hands of the journal's editor and reviewers. Such a pre-reviewer should be thanked with an acknowledgment, usually just before the references in the manuscript

Revision of Manuscript

Every good manuscript goes through several revisions before it is finally submitted for publication. How many times this will have to be done to produce the clearest, most concise final product will vary with the complexity of the project, the experience of the authors, and the degree to which they disagree on revisions. A good paper is not ready for submission the minute its parts have been assembled for the first time. Manuscripts sent to an editor in this condition have little chance for acceptance.

Preparation and Submission of Final Draft

Prospective authors must read carefully the "instructions for manuscript preparation" section of a recent issue of the journal to which the paper will be sent. Then they must follow these instructions to the

letter. Following detailed submission instructions may seem a tiresome chore in the euphoria that comes with completion of a project, but it is no less important than the other stages of manuscript preparation.

Each author should receive a copy of the final manuscript as it is to be submitted for publication. In addition to ensuring that he or she agrees with everything that is said in the paper, this will guarantee that each author's name and professional affiliations are shown correctly on the title page. First-time authors should keep in mind that a published paper is a permanent, public record.

STRUCTURE OF A CASE REPORT

Like other forms of scientific writing, the case report should be organized so that its message is presented to the reader in the clearest, most logical fashion. Properly written case reports, therefore, "are not just baskets carrying unconnected facts, like the telephone directory; they are instruments of persuasion. In keeping with this notion that an author needs to convince the reader of the validity of what is said, here is a logical arrangement for any scientific paper that applies to case reports:

1. Statement of problem or posing of question
2. Presentation of evidence
3. Explanation of the validity of the evidence presented
4. Implications of the evidence: initial conclusion or answer s. Statement of additional supporting evidence
5. Assessment of conflicting evidence
6. Final conclusion.

Case reports are generally brief, and need not have separate sections for each of these components. The great majority have the logical format of Introduction, Case Summary, Presentation of Additional Data (if appropriate), Discussion, and References. Some journals also require a brief abstract preceding the introduction. However, the fundamental purpose of the paper is to communicate information, and some cases may lend themselves better to some variation on this plan. In all cases, however, the "critical argument" sequence outlined above should be considered and its elements addressed.

Most case reports will lend themselves best to the structure summarized in the next sections.

Introduction

The introduction announces to the reader what the paper is about and why this is important. It should not generally be more than a paragraph. If definitions or a description of the syndrome being presented are necessary for the reader's understanding of the case summary , then the introduction should provide these.

Case Summary

This is the heart of the case report, and must include enough data to convince a critical reader that the author's contention is correct. Only those aspects of the patient's illness, the event, or the procedure described that are crucially important for this purpose should be included. Inexperienced authors often have difficulty distinguishing the clinically relevant from the extraneous in this respect, and having the case reviewed by an expert can be helpful.

Although only elements important to the report should be provided, the case summary should still adhere to the traditional format for presenting a case:

1. Chief complaint
2. Current illness
3. Personal and family history
4. Occupational and environmental history
5. Physical examination findings
6. Initial laboratory findings
7. Initial x-ray findings
8. Working diagnosis
9. Hospital course
    a. Initial treatment
    b. Chronology of response to treatment; complications; new forms of treatment; later physical or laboratory findings
    c. Current status or outcome; revised diagnosis or problem list.

In most cases, some of these components will not be needed, and others need only be mentioned briefly, but the sequence should be followed.

Tables and Illustrations

These are really part of the case summary, unless the paper also contains a "Results" section following the case summary that presents data from original research based on or developed from the patient's illness. Numerical data such as long sequences of arterial blood gas values or other measurements, if included at all, should generally be provided in a table. An alternative is to display these data visually using a graph, particularly when a small number of functions (such as PO_2, PCO_2, and pH) is reported repeatedly to describe the course of the patient's condition. Most journals publish explicit instructions for the preparation of tables and figures in their "instructions for authors," and these must be followed.

Radiographs and other images are often reproduced in case reports because of their important role in diagnosing and managing diseases of the chest. However, it is difficult to reproduce chest films in a manner that shows the desired features clearly to the reader of the final published product. If the chest x-ray was normal, or showed right upper lobe consolidation or a pneumothorax, this can simply be stated. If information vital to the case report is on film, appropriate consultation with both the radiologist and a medical photographer will ensure the best possible reproduction. Once the radiologist has pointed out exactly what and where the features are that should be illustrated, you should go over this with the photographer so that he or she can use developing and printing techniques that best show these features.

Discussion

A case report should teach, and the teaching is done in the discussion. The discussion amplifies the case summary by pointing out and explaining its unique or otherwise important aspects. The elements of "critical argument" listed at the beginning of this section should be included. Typically, a case report's discussion section accomplishes this in three or four paragraphs that contain the following:

- an initial, brief summarizing statement of the reason for reporting the case;
- a concise description of the disease or condition illustrated, if this was not done earlier, including any important features not found in the case presented;
- a brief discussion of additional evidence, alternative explanations, atypical or complicating features, or other factors that need to be mentioned;
- a clear statement of the lesson(s) to be learned from the case.

Although some journals accept lengthy literature reviews introduced by case reports, the majority insist on brevity. Virtually all published case reports have shorter discussion sections than originally submitted to the editor. Thus, the author's goal should be to make only the points that have to be made, do this with the fewest possible words, and eliminate everything else. A five-page manuscript is no less a contribution than one four times that length, and it will be received more appreciatively by the journal's editor.

References

The references at the end of the paper should consist of a small number of citations. They should not be merely a random sampling of previously published work on the subject of the report, but should serve two specific purposes. First, because lengthy descriptions of the disease, treatment, etc. cannot be included in a case report, you should cite one or two authoritative sources of more complete information. These could be recent review articles or chapters in authoritative textbooks. Second, the references should back up any specific or controversial points made in the report. Citations should be specific, providing the reader with the exact source for the information referred to, rather than, for example, to an entire book, and they should also be to reference sources that are readily available to the reader of the paper. You must not cite references found in other papers unless you have personally read them; cited articles do not always contain the information suggested by their titles, or data justifying the conclusions drawn by previous authors who cited them

AVOIDING COMMON MISTAKES IN CASE REPORT WRITING

Table 16-3 and summarizes 12 mistakes commonly made by authors of case reports, particularly those with little previous experience with such projects.

Table 16.3. Twelve common mistakes made by authors of case reports.

In Selecting the Case Itself

1. Tunnel vision: Unfamiliarity with experience or practice outside of one's own geographic region.
2. Insufficient documentation of the case: Inadequate data base to establish the diagnosis or feature being reported.
3. Insufficient documentation of intervention or outcome: Failure to prove the value of a new device, technique, or therapy.
4. Poor patient care: Documentation of results of bad clinical practice rather than spontaneous events.
5. Erroneous premise: Mistaken physiology or inadvisable therapy that happened to be associated with a favorable outcome on this occasion.

In Preparing the Manuscript

6. Wrong journal: Inappropriate subject matter or format.
7. Literary inexperience: Unfamiliarity with scientific, professional, or medical writing.
8. Inadequate literature review: Failure to locate and cite important background material.
9. Ineffective illustrations or tables: Poor selection and communication of visual or numerical data.
10. Poor references: Inappropriate selection of documentation or sources of more extensive information
11. Technical mistakes: Failure to adhere to journal's published instructions for manuscript preparation; inadequate proofreading prior to submission.
12. Non-revision: Failure to follow through with suggestions by reviewers and editor to resubmit revised manuscript.

Tunnel Vision

Through no fault of their own, prospective authors of case reports may believe their observation to be unique and "reportable" because they and those around them have never heard of it before, when in fact it is a well-documented but infrequent phenomenon. Steps 2 and 3 in Table 16-2, review of the literature and discussion of the case with an "expert," will prevent expenditure of needless effort in developing the project further. Another form of tunnel vision is to assume the rest of the world uses the same procedures and equipment as you do. Stating that the patient underwent "standard tests for collagen vascular disease" will confuse many readers (What tests? Which collagen vascular diseases?) and fails to accomplish the basic goal of all scientific writing, which is to communicate information accurately and completely.

Insufficient Documentation of Case

One of the most frequent reasons that editors reject case reports submitted to their journals for publication is an inadequate database. It is not enough that the author really, truly believes that the

patient had acute respiratory distress syndrome (ARDS) secondary to legionnaire's disease. This diagnosis must be rigorously proven to the editor and to the reader. The proof established not only with the accepted definitions and criteria but also by careful exclusion of other possible diagnoses. Case reports often provide more extensive documentation than would ordinarily be obtained in clinical practice. If features of the case are atypical or do not follow the expected course for the condition being reported, these must be explained convincingly

Insufficient Documentation of Intervention

The author's belief that treatment of condition A with intervention B resulted in outcome C may be the reason for submitting the paper to a journal. Whether the journal's editor accepts that reasoning for publication is likely to depend upon the means by which the author proves his or her case. If a patient with ARDS recovers after receiving large doses of corticosteroids, one is tempted to conclude that recovery is related to the drugs. However, the essence of the scientific method is that such associations can, and must, be established far more rigorously. Perhaps the best way to begin is to consult with an individual having relevant experience.

Poor Patient Care

Some complications or outcomes submitted to journals as case reports occur not as spontaneous events but as a result of bad clinical management. Unless the intention of the author is to emphasize this for teaching purposes, such a paper would not be accepted for publication. Doing so would appear to give the journal's approval of the patient care described.

Erroneous Premise

Case reports that are based on a faulty understanding of physiology are occasionally submitted for publication (and some have even been published). Others document potentially dangerous interventions that happened to be followed by a favorable outcome. In such cases, the problem is usually inexperience on the part of the would-be author, failure to review the relevant literature, and lack of discussion and consultation with more experienced individuals.

Submission to the Wrong Journal

Medical journals are published for well-defined audiences. Each has its own requirements for the format, length, and subject matter of papers it publishes. Prospective authors must familiarize themselves with the journal to which they intend to submit their manuscript. They must read carefully the journal's "instructions for authors," and also read the articles that journal publishes. If the author is unfamiliar with a particular journal, an appropriate question would be why he or she wishes to submit a paper to it. If you do not read the journal, very likely your peers do not read it either.

Literary Inexperience

Often, authors submitting a case report have never before attempted to write something for publication. Few health care practitioners have had formal training in scientific writing and manuscript preparation. They should not be dissuaded from trying, because a case report is the ideal first paper, but will need to study the subject or to collaborate with someone with appropriate experience.

Inadequate Literature Review

To serve its purpose of contributing to the sum of our knowledge on a particular subject, the case report must place the condition or event described in appropriate context. This requires that the author become thoroughly familiar with present knowledge on the topic prior to beginning to write. Cited references should be recent, and many more sources should be read than cited. The author must be confident that others have not published similar reports in the journal or field involved. Although it is a minor literary effort in comparison to some other types of publication, the case report is nonetheless a great deal of work-at least as much as a college term paper. Reviewing the literature comprises much of this work, but without this part of the project, the whole cannot succeed.

Ineffective Illustrations or Tables

Illustrations and tables are crucial to the case report's effectiveness, as already discussed, and must be of professional quality. Photographs of x-rays should not show more than the feature of interest: an entire 14- by 17-inch chest film should not be reproduced if a stenotic area in the cervical trachea is being illustrated. The film should be photographed and printed so as to emphasize the desired feature, and arrows should be included if the point referred to is not obvious to the reader. Tables are better than inclusion within the text if a series of numerical values must be reported, although such data have to be trimmed to only those results that are crucial to the case. Even if 16 different parameters of the patient's pulmonary function were measured, these should be omitted unless the value of the case rests on them.

Poor References

An author provides references so that the reader can amplify and verify statements made in the paper. All cited sources must therefore be recent and accessible, in frequently used books or journals. The original description of an illness or therapy might be an exception to this rule, but review articles and chapters in textbooks must not be.

Technical Mistakes

There is no excuse for submitting a manuscript that does not conform to the journal's published instructions for authors. This applies not only to the paper's length and format, but also to restrictions on tables and illustrations and to the citation of references. In their rush to get the final draft into the mail, authors often fail to proofread their manuscript carefully enough, and send it to the journal's editor with typographical errors, punctuation mistakes, and misspelled words. Every reference must be double-checked against its original source (not the rough draft), and, if necessary, someone not connected with the paper should proofread it for grammar and typing.

Failure to Revise the Manuscript after Editorial Review

Not one manuscript in 100 is accepted on its first submission to a peer-review journal without requests that the author make some revisions. However, many papers fail to reach publication because the modifications suggested by reviewers or editor are not followed up by the author. If the manuscript cannot be made acceptable for publication, it will be rejected; otherwise, it will come back to the author with a list of changes the editor considers important in order to make it acceptable. Occasionally these suggestions are unreasonable or even impossible. For example, when additional data are required and the patient is no longer available or the records are lost-but in most cases the author could make the changes within a few hours. Comments from reviewers and editors are intended to improve the paper and insure that it does not mislead or contain erroneous information. If the project was worth pursuing in

the first place, it is worth the effort to revise the manuscript and attempt to meet the criticisms and suggestions made for it.

QUESTIONS

True or False

1. A case report consists of a description of an individual illness, or treatment, followed by a discussion of the unique features it demonstrates.
2. A previously unreported clinical feature of a known disease is a reportable as a case study.
3. A new diagnostic test or other assessment is not an appropriate report topic even if it is illustrated by an individual case.
4. Unlike an original study, the case report does not require a literature review.
5. A logical arrangement of the case report is: (1) statement of problem, (2) presentation of evidence, (3) explanation of validity of evidence, (3) implications of evidence, (4) assessment of conflicting evidence, (5) conclusions.
6. One common mistake made by authors of case reports is tunnel vision, or unfamiliarity with medical practice outside of one's own geographic region.

Chapter 17. The Poster Presentation

In the chapter on writing the abstract, we discussed the fact that scientific presentations at most medical conventions are either lectures or poster presentations. Poster sessions allow direct communication with the author(s) in a more relaxed and informal atmosphere and more time for the audience to examine illustrations. These factors help to ensure proper interpretations of the information presented. The author(s) benefit too, in that dialogue with other researchers with similar interests may stimulate new ideas.

There are two main formats for poster presentations. In the first, or *Open Session,* posters relating to various topics are set up in a large exhibit hall. The author(s) stand next to these posters and answer questions as interested viewers come by at random over a period of two to three hours. A more formal approach is the *Poster Symposium.* Posters of related subject matter are grouped in a specified area at a specific time. A moderator, chosen for his or her expertise in the particular subject, directs each session. A relatively small audience views posters for a short period before the session begins. Then, at the moderator's request, the author gives a short (one to two minutes) oral presentation outlining the significance of the study. Following this, the audience can discuss the work with the author. Often, discussions are generated among the members of the audience under the guidance of the moderator. Compared to oral abstract presentations, poster presentatins are less formal, less anxiety producing for the beginner, and may have more of the atmosphere of a social event.

Poster quality at scientific meetings is sometimes barely acceptable. This is due primarily to poor planning and inexperience. The following suggestions should help to improve the effectiveness of a poster presentation.

LAYOUT

A bulletin board will be provided by the sponsor of the conference for mounting the poster with thumbtacks. The usable surface of this board is four feet high and six feet long and is elevated about two to three feet from the ground. The information presented by the poster follows the same format as the abstract of an article. The poster format includes a title (along with the names of the authors, institution, city, and state) introduction, methods, results, and conclusion (Fig. 17-1).

PLANNING

Begin by drawing a rough sketch. Try various styles of data presentation to achieve *clarity* and *simplicity.* This sketch can later be used to help set up the poster at the time of presentation. Decide on the length and content of the text, avoiding acronyms, abbreviations, and jargon. The poster should be self-explanatory. Avoid the temptation to overload the poster with too much information. Keep in mind the major ideas to be communicated. Creating the text of the poster is similar to writing an abstract. In fact, to have a poster presentation accepted by the program committee, one first has to submit it in abstract form. Some authors give out copies of this abstract at the poster session.

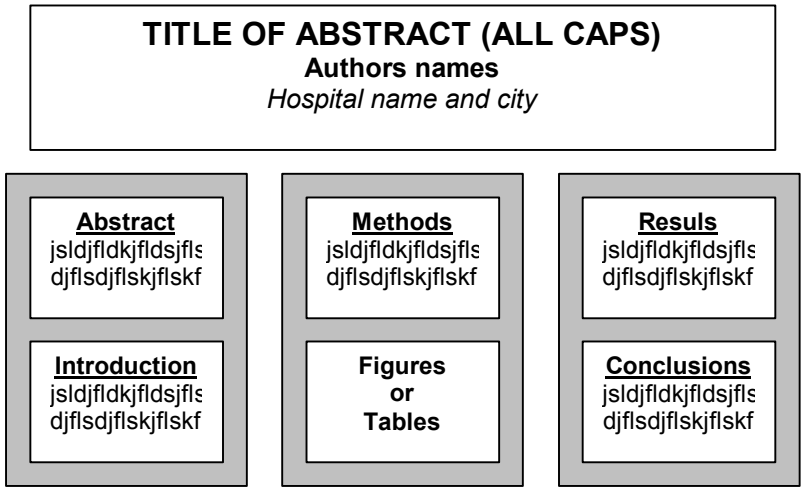


Figure 17-1. Example layout of a poster presentation.

The organization of information (eye movement) should flow naturally down columns or horizontally along rows. Material at the top and center of the poster will attract the most attention. Material in the lower left and right corners is less important, and these areas may be left blank to provide visual rest and to highlight the main points. Once the rough sketch is completed, it is helpful to redraw it full size on a blackboard. This will help to establish proportions and sizes of each section. Indicate text with horizontal lines, and draw in rough tables and figures.

MATERIALS

There are two basic approaches to making the poster itself. One way is to hire a graphic artist to do the layout and then print the poster on one large piece of paper or plastic laminated light cardboard. This approach usually results in a very good looking poster with beautiful color coordination and lettering. However, it is quite expensive; you can expect to pay several hundred dollars.

The other approach is do-it-yourself. I have found the best design is modular (i.e., text and illustrations mounted on cardboard) for convenient transportation. This modular approach will also allow one to reuse portions of the poster (e.g., subtitles like Introduction, Methods, and so on). Go to a graphic arts store and buy several pieces of poster board with the desired background color (gray or hunter green are good). Have the store cut the board in half (ie, into pieces measuring about 16" x 20"). Next, lay out your sections on regular 8 ½" by 11" paper in landscape view (sideways). Use 20 or 22 point Arial font for the body and 48 point Arial Black font for the titles (ie, Introduction, Methods, etc). These are just guidelines, adjust as necessary. Use removable spray adhesive (also from the graphic arts store) on the back of the printed sections. Then carefully place them on the poster board, two per board (see Figure 17-2). Finally, print out a banner on a single sheet of paper, again in landscape view, using 36 point Arial Black for the title, 22 point Arial for the authors names under the title, and 20 point Arial italic under that for the names of your hospital. Of course, this is too small for the actual banner. Go to a printer (eg, Kinkos) and have them enlarge it with a banner photocopy machine until it is about 3 ½ to 4 feet long and about 9" wide. You will get better quality if you have a graphic arts store (or your hospital's print shop) make the banner for you, but it will cost more.

After each of the components of the poster has been completed, lay them out on the floor to be sure they fit into the 4 x 6 foot space limitation. This gives the presenter a chance to make corrections, and will give an idea of how to set up the actual presentation on the poster board. Once satisfactory results have been achieved, stack the sections of the poster together and enclose them in a protective cardboard cover. If one is traveling to the conference by plane, take the package on board to avoid damage or loss. Make sure that it is small enough to fit under the seat or in the overhead cargo areas. Be sure to bring clear or white thumbtacks. The components should not be mounted on the poster boards with staples or nails. The poster should be assembled at least 15 minutes before the poster session begins. The presenter will be responsible for removing it afterwards.

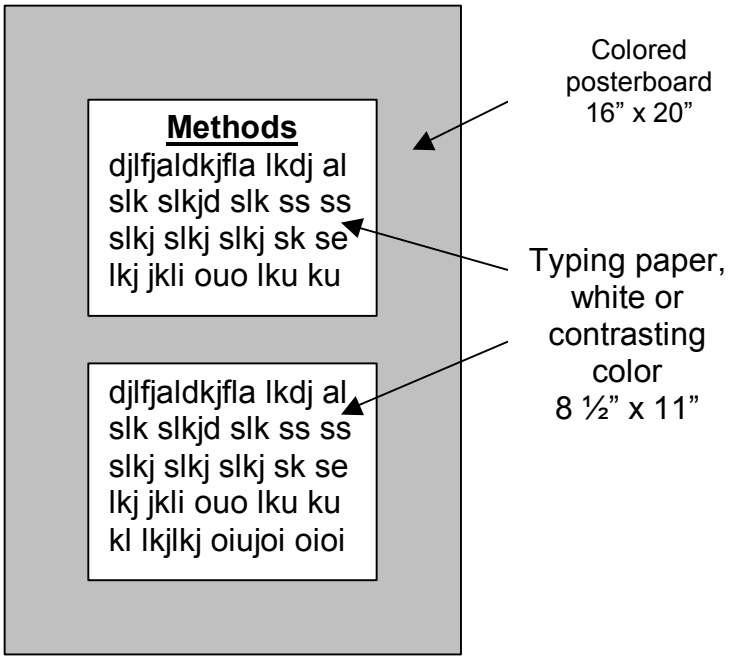


Figure 17-2. Layout of an individual component of the poster presenation.

QUESTIONS

True or False

1. Poster presentations allow direct communication with the author in an informal atmosphere and more time for the audience to examine illustrations.
2. Posters are usually not accepted in scientific meetings unless they are created by professional graphic artists.
3. Poster layouts are similar to that of abstracts and papers: Title, Introduction, Methods, Results, and Conclusions.
4. You should strive for clarity and simplicity in your poster layout.

SECTION VI APPENDICES

Appendix A. Glossary

This is a "teaching" glossary. It contains many terms and explanations that do not appear in the text.

Accuracy: The degree to which a measured value reflects the true value; sometimes expressed as the difference between true and measured values as a percentage of the true value (see Error).

Algorithm: A well defined set of rules which, when routinely applied, lead to a solution of a particular problem (often expressed as a flowchart).

Alpha (α) level: The maximum tolerable probability of making a Type I error. (see significance level)

Analysis of variance (ANOVA): An inferential statistical test used to test the significance of differences between three or more mean values.

Bell curve: The characteristic shape of the normal and *t* distributions; having the shape of a vertical cross section of a bell.

Beta (β) level: The probability of making a Type II error.

Benchmarking: A procedure for identifying best practices among a group of similar organizations.

Bias: Deviation of results or inferences from the truth, or processes leading to such deviation; the difference between the mean value of a set of measurements and the true value.

> *Ascertainment bias* arises when a higher exposure to a risk factor causes a higher probability of detecting the event of interest.
>
> *Detection bias* is the tendency to look more carefully for an outcome in one of two groups being compared.
>
> *Publication bias* occurs when the publication of research depends on the direction of the study results and whether they are statistically significant.
>
> *Recall bias* occurs when patients who experience an adverse outcome have a different likelihood of recalling an exposure than the patients who do not have an adverse outcome, independent of the true extent of exposure.
>
> *Selection bias* may occur whenever a treatment is chosen by the individual involved. A classic example of this problem occurred in the Lanarkshire milk supplementation experiment of the 1920s. In this trial, 10,000 children were given free milk and a similar number received none. The groups were formed by random allocation. Unfortunately, well-intentioned teachers decided that the poorest children should be given priority for free milk rather than sticking strictly to the original groups. The consequence was that the effects of milk supplementation were indistinguishable from the effects of poverty. Another example of selection bias is *volunteerism*, where the study entrant selects himself.

Blind (or Blinded or Masked): The participant in a study is unaware of whether patients have been assigned to the experimental or control group. Anyone associated with the study (patients, researchers, data analysts, writers, etc.) may be blinded. To avoid confusion, the term *masked* is preferred in studies in which vision loss of patients is an outcome of interest. If both the researcher and the patient are

unaware of the assigned treatment, the study is *double blinded.* If only one of the two is blinded, the study is *single blinded.*

Blind assessment first began in the late 18th century as a tool for fraud detection mounted by elite mainstream scientists and physicians to challenge the suspected delusions or charlatanism of unconventional medicine. Some of the first experiments were carried out to evaluate mesmerism, and were literally conducted with blindfolds. They took place in France at the house Benjamin Franklin, the American minister plenipotentiary, who was head of a commission of inquiry appointed by King Louis XVI.

Calibration: The adjustment of the output of a measuring device to match the value of a known input. A very common example in respiratory care is the calibration of an oxygen analyzer by exposing the sensor to pure oxygen and adjusting the readout to display 100%.

Calibration verification: Measuring a known (assumed true) value with a calibrated device and noting the difference between the measured and true value. If the difference is below some predetermined threshold, the device is judged accurate enough for use. If not, it is recalibrated.

Causation: The process whereby a given event, called a cause, *invariably* precedes a certain other event, called the effect.

Clinical trial: Planned experiment involving participants, usually patients, to determine the most appropriate therapy by comparing one therapeutic approach with another, usually standard care.

Confounding factor (variable): A factor that distorts the true relationship among the study outcome variables because it is also related to the outcome variables. Confounding variables are often unequally distributed among the groups being compared. Randomized studies are less likely to have their results distorted by confounding factors than are non-randomized studies.

Continuous variable: A variable that can theoretically take any value and in practice can take a large number of values with small differences among them.

Control: Any operation that is designed to limit any of the conceivable sources of error (confounding factors) in a study. Experimental control refers to the manipulation of conditions under which the observations are made. Statistical control involves the treatment of data to remove the effects of confounding factors.

Control group: Subjects who are as closely as possible equivalent to an experimental group and exposed to all the conditions of the study except the experimental treatment. In many studies, the control group receives either the standard of care or no treatment at all.

Cost analysis: If two strategies are analyzed but only costs are compared, this comparison would address only the resource-use half of the decision (the other half being the expected outcomes) and is termed a cost analysis.

Cost benefit analysis: A form of decision analysis in which both the costs and the expected outcomes (benefits) are measured on the same continuous scale (either both as costs or both as dimensionless numbers) and compared as a ratio (either as cost/benefit or as benefit/cost). This relationship allows the comparison of two or more courses of action (e.g., new therapies or purchases of equipment).

Cost-effectiveness analysis: An economic analysis in which the expected outcomes are expressed in natural units. Some examples would include cost per life saved or cost per unit of blood pressure lowered.

Cost minimization analysis: An economic analysis conducted in situations where the expected outcomes of two or more alternative courses of action are identical, so the only issue is their relative costs.

Cost-utility analysis: A type of cost-effectiveness analysis in which the expected outcomes are expressed in terms of life-years adjusted by peoples' preferences. Typically, one considers the incremental cost per incremental gain in quality adjusted life-years (QALYs).

Crossover design: The same individual is included in both treatment and control conditions or in multiple treatment conditions in the same study; the individual acts as his own control. Random allocation is used to determine the order in which the treatments are received. The simplest design involves two groups of subjects. One group receives each of two treatments, A and B, in the order AB, while the other group receives them in the reverse order, BA. Because the treatment comparison is within subject rather than between subjects, fewer subjects may be required to achieve a given statistical power. Carryover effects may be a problem. An attempt to minimize this problem is to include a wash-out period between the two treatments (the period required for a treatment to cease to act once it has been discontinued).

Crossover effects: Residual effects of the treatment received on the first occasion that remain present into the second occasion.

Decision analysis: A systematic approach to decision making under conditions of uncertainty. It involves identifying all practical alternatives and estimating the probabilities of potential outcomes associated with each alternative, assigning value weights to each outcome, and on the basis of probabilities and values, arriving at a quantitative estimate of the relative merit of the alternatives.

Deductive reasoning: Reasoning that proceeds from a general law to a conclusion that is specific to a particular situation. Reasoning from general to particular.

Degrees of freedom: An elusive concept that occurs throughout statistics. Essentially, it means the number of independent units in a sample used to calculate a statistic. Another way of thinking about it is 1 minus the number of sample values that can be determined from knowledge of the other values. For example, if the mean value of 3 numbers is 4 with $x_1 = 3$ and $x_2 = 4$, then x_3 must equal 5. One of the three sample values can be determined by knowledge of the other two, so the "degrees of freedom" is $n - 1 = 2$.

Delphi technique: A questionnaire strategy in which a panel of experts complete consecutive questionnaires generated on the basis of the answers to the previous questionnaires; designed to achieve a consensus among the panel of experts.

Dependent variables: The variables the investigator measures in response to the causal (treatment or independent) variable; outcome variables.

Dichotomous variable: A variable that can take one of two values, such as yes or no, dead or alive, 1 or 0, etc.

Drift: The gradual changing of an instrument's accuracy during use due to uncontrollable electrical, chemical, or physical factors.

Effect size: The difference in outcomes between the intervention and the control groups divided by some measure of variability, usually the standard deviation.

Effectiveness study: A study designed to answer a question of the type: " Does the treatment work in clinical practice settings with unselected patients, typical care providers, and usual procedures?"

Efficacy study: A study designed to answer a question of the type: "Does the treatment work in a tertiary care setting with carefully selected patients under tightly controlled conditions?"

Endpoint: Endpoints are events our outcomes that lead to completion or termination of a study or follow-up of an individual (e.g., death or major morbidity).

Error: The difference between measured and true values, sometimes expressed as a percentage of the true value; bias plus imprecision.

Experimental variable: The variable(s) manipulated by the researcher; the independent variable(s).

Exposure: A condition patients come in contact with (either a potentially harmful agent or a potentially beneficial one) that may affect their health.

False negative: In a treatment study, treatment is considered ineffective when it actually is effective. In a diagnosis study, the patient has the target condition, but the test suggests the patient does not.

False positive: In a treatment study, treatment is considered effective when it actually is not effective. In a diagnosis study, the patient does not have the target condition, but the test suggests the patient does.

Gold standard: A method having established or widely accepted accuracy, providing a standard to which measurements can be compared (eg, for calibration).

Generalizatility: The ability to apply results from a sample to a population. For example, results from a study of brohchodilators on a group of asthmatics may be generalizable to all asthmatics in the world if the sample had the same characteristics as the world population of asthmatics and in the same proportion (e.g., age, gender, severity of illness, socioeconomic status, etc).

Halo effect: The tendency to overrate a subject's performance on some task because of the observer's perception of the subject doing well during a prior evaluation.

Hawthorne effect: A term used for the effect that might be produced in an experiment simply from the awareness by the subjects that they are participating in some sort of scientific research (due to the novelty of the situation of being treated in a special way). The psychological reaction to the study conditions can be mistaken for the effect of the experimental variable(s). It was first described in some classic experiments at the Hawthorne plant of the Western Electric Company in Chicago during the late 1920s and early 1930s.

Hello-goodbye effect: A phenomenon first described in psychological research, but on which may arise whenever a subject is assessed on two occasions, with some intervention between the visits. Before an investigation, a person may present himself in as bad a light as possible, thereby hoping to qualify for treatment, and impressing staff with the seriousness of his problems. At the end of the study, the person may want to please the staff with his improvement, and so may minimize any problems. This minimization may lead to the false conclusion that some improvement exists when none has actually occurred, or to magnify the beneficial effects that did occur.

Hidden time effects: Effects that arise in data sets that may simply be a result of collecting the observations over a period of time.

Historical controls: A group of patients treated in the past with a standard therapy, used as the control group for evaluating new treatment on current patients. Although used frequently in medical

investigations, this study design is not recommended because possible biases (due to other factors that may have changed over time) can never be eliminated.

Hypothesis: A statement of relationship among variables that is assumed true until data are obtained that prove it wrong. The *null hypothesis* is usually a statement of no difference or no association between groups (the observed difference in values of a statistic are due to chance alone). The *alternate hypothesis* is the opposite of the null and usually states that there is a difference or that the difference is in a particular direction (one group statistic is larger or smaller than another).

Imprecision: The extent to which repeated measurements of the same quantity vary from one another; usually characterized by the variance or the standard deviation as an estimate of the random error of measurements.

Incidence: The number of new cases of disease occurring during a specified period of time; expressed as a percentage of the number of people at risk.

Inclusion criteria: Criteria used to define the population who will be eligible for a study.

Independent variables: Explanatory or predictor variables that may be associated with a particular outcome.

Induction: A process of reasoning that proceeds from specific facts in a particular situation to general propositions or laws. Reasoning from particular to general.

Inference: A judgment based on other information rather than on direct observation. Statistical inference is the process by which one is able to make generalizations from the data.

Informed consent: The situation where a competent person, in possession of all the relevant facts, has agreed to participate in a research study.

Intention to treat principle (analysis): A procedure in which all patients randomly assigned to a treatment are analyzed together as representing that treatment, whether or not they completed or even received it. This method is intended to prevent bias due to using compliance with treatment (a factor often related to outcome) to determine the groups for comparison. It maintains the power of randomization.

Interaction effect: The effect of two or more independent variables acting in combination rather than independently of one another; the effect of one factor is not consistent over al levels of another individual factor in a two factor (two way) analysis of variance (ANOVA).

Interval level: Measurement scale which has the following properties (a) the values are distinguishable (b) they are ordered (c) the intervals between the points on the scale are equal (d) the zero point is not absolute; it does not represent the absence of the quantity being measured. An example is temperature on the Celsius or Fahrenheit scales.

Inverse rule of 3s: A rough rule of thumb that says if an event occurs on average once every "*n*" days, we need to observe 3*n* days to be 95% confident of observing at least one event. For example, if the event occurs once every 10 days, we need to observe 3 x 10 = 30 days for at least one event to occur.

Likert-type scale: Scales, typically with 3 to 9 possible values, that include extremes of attitudes or feelings (such as strongly agree, agree, undecided, disagree, strongly disagree). A number is attached to each possible response and the sum of these ratings is used as the composite score. These scales are used to obtain ratings from study participants. A commonly used Likert-type scale is the Apgar score used to evaluate the status of newborn infants.

Linearity: The degree to which variation in the output of an instrument follows input variation.

Literature review: A summary of earlier published work on a topic of interest, containing a critical review of what is known.

Longitudinal study: Study that requires obtaining repeated measure over time.

Meta-analysis: A quantitative procedure for combining the results of several studies into a pooled estimate of the independent variable on the dependent variable.

N of 1 clinical trial: A special case of a crossover design for determining the efficacy of a treatment for a specific patient. The patient is repeatedly given a treatment and placebo, or different treatments, in successive time periods.

Nonparametric test: A class of statistical test that is not based on the estimation of a parameter for the underlying population, usually does not assume a particular sampling distribution, and can be used with nominal or ordinal data.

Negative results: Research findings that suggest the null hypothesis be accepted; non-significant results.

Nominal variable: Variable that takes on distinct, mutually exclusive categorical values (e.g., male of female, yes or no). The values have no particular order and the intervals between the values are meaningless. Numbers associated with the categories have no quantitative value and are used only as labels.

Ordinal variable: A variable having the following properties (a) mutually exclusive, (b) ordered but the intervals between the values are not equidistant, (c) there is no meaningful zero point.

Observational study: Study providing questions about overt behaviors or events and answers using human observers to record the behaviors or events over a period of time. The researcher has little or no control over the behaviors or events.

Occam's (Ockam's) razor: Also known as the parsimony principle, popularized by William Occam, a 14^{th} century philosopher. The principle states that unverified assumptions should be kept to the bare minimum and the hypothesis with the fewest assumptions is preferred. Michael Faraday warned against the tendency of the mind "to rest on an assumption" and when it appears to fit in with other knowledge to forget that it has not been proved.

Outlier: An observation that appears to deviate markedly from the other members of the sample. Extreme values may reflect some abnormality in the measurement process.

Parameter: A numerical characteristic of a population (e.g., the population mean) or a mathematical model (e.g., the probability of success in binomial distribution).

Percentile: The point in a cumulative percentage plot below which the percent of cases indicated by the given percentile falls. For example, an IQ of 100 represents the 50^{th} percentile of intelligence scores; 50% of people have IQ scores below 100. In a normal distribution, the mean represents the 50% percentile.

Phase I drug study: Studies that investigate a drug's physiological effect or ensure that the drug is not toxic, often conducted with normal volunteers.

Phase II drug study: Initial studies on patients, which provide preliminary evidence of possible drug effectiveness.

Phase III drug study: Randomized control trials designed to definitively establish the magnitude of drug effects.

Phase IV or post-marketing surveillance drug study: Studies conducted after the effectiveness of a drug has been established and the drug marketed, typically to establish the frequency of unusual toxic effects.

Pilot study: A study carried out before a research design is completely formulated. The purposes of a pilot study are to assist in (a) the definition of the problem, (b) the development of the hypotheses, (c) the estimation of sample sizes and, (d) the establishment of priorities for further research.

Placebo: (From the Latin "I will please"). A treatment designed to appear exactly like a comparison treatment, but which has no active component or effect. It is used to control for psychological bias (see Hawthorne effect) by matching the experimental and control groups in terms of equivalent exposure to the treatment administration. In surgical studies of surgical procedures, the placebo treatment is called a sham surgery (e.g., opening the abdomen but not removing the appendix).

Play-the-winner rule: A procedure sometimes considered in clinical trials in which the response to treatment can be classified as positive (a success) or negative (a failure). One of two treatments is selected at random and used on the first patient. If the response is positive, the same treatment is used on the next patient; if the response is negative, the other treatment is used on the next patient. The goal is to minimize the number of patients assigned to the inferior treatment.

Power: The probability of rejecting the null hypothesis when it is false; correctly identifying an effect when it is there. Power = $1 - \beta$. When comparing tests of the same hypothesis, the one with the highest power is preferred. Power is also the basis for estimating the sample size needed to detect an effect of a particular magnitude (effect size).

Precision: The degree to which repeated measurements of the same quantity agree with each other. Often quantified in terms of the variance or standard deviation as an estimate of the random error of measurements. (see imprecision).

Prevalence: A measure of the number of people in a population who have a particular disease at a particular time. Point prevalence is the number of cases at a particular moment divided by the number in the population at that moment. Period prevalence is the number of cases during a specified period divided by the number in the population at the midpoint of the period.

Prospective study: A type of research design, also known as a longitudinal study, in which the collection of data proceeds forward in time (as opposed to a retrospective study). A cohort study is a type of prospective study in which the same group of subjects is followed over a period of years and data are periodically recorded.

P value: The probability that the observed results or results more extreme would occur if the null hypothesis were true and the experiment were repeated many times.

Quality-adjusted life-year (QALY): A unit of measurement for survival that accounts for the effects of suboptimal health status and the resulting limitations in quality of life. For example, if a patient lives for 10 years and her quality of life (measured on some scale) is decreased by 50% because of chronic lung disease, her survival would be equivalent to five quality-adjusted life-years.

Random: Process governed by chance such that the occurrence of previous events is of no value in predicting future events. For example, the long term probability of observing "heads" on the toss of a coin is 0.5, but the actual outcome of the next toss cannot be determined based on the number of heads

in previous tosses. The "gambler's fallacy" occurs when, for example, a gambler observes 10 heads in a row and then bets heavily that the next toss will not be heads because the probability of 11 heads in a row is very low. In fact, 11 heads in a row is rare but that fact is irrelevant; each toss is independent of the last with probability of heads on a given toss remaining at 0.50.

Randomization: A technique in experimental research to equalize the composition of the various groups under study so that they are as similar as possible on all pertinent characteristics (including confounding factors). Subjects are allocated to the different study groups according to the laws of chance (e.g., by flipping a coin, drawing ballots, or using a table of random numbers).

Ranking: The process of sorting a set of values into either ascending or descending order.

Ratio level: Measurement scale which has the following properties (a) the values are distinguishable (b) they are ordered (c) the intervals between the points on the scale are equal (d) the zero point is absolute; it represents the absence of the quantity being measured. Examples are height, weight, and temperature measured on the Kelvin scale.

Regression to the mean: The phenomenon first noted by Sir Francis Galton (one of the originators of statistics) that "each peculiarity in man is shared by his kinsmen, but on the average to a less degree." For example, tall parents tend to produce offspring who are tall, but on average, are shorter than their parents. The term is used now to describe how a measurement that appears extreme on its first measurement will tend to be closer to the average value for a later measurement. For example, in a screening program for hypertension, only persons with high blood pressure are asked to return for a second measure. But on average, the second measure will be less than the first.

Repeatability: The closeness of repeated measurements of the same quantity by the same observer under the same conditions, and using the same equipment over relatively short time intervals; precision.

Reproducibility: The closeness of repeated measurements of the same quantity by different observers under different conditions and different equipment over a relatively long time.

Retrospective study: The type of design in which the dependent variable is observed first and the data are traced back and related to possible relevant independent variables that are hypothesized as being associated with the dependent variable; considered to be a non-experimental design.

Rosenthal effect: The phenomenon where the expectations of the researchers in a study influence the outcome. For example, if an observer believes that a particular treatment is effective he may under-report observations inconsistent with this belief.

Sampling: The process of selecting a fraction of a population for inclusion in a study.

> *area sampling:* A probability sample in which the primary sampling units are households used in large-scale studies conducted by the government. On a smaller scale, beds in a hospital could serve as a sampling unit.
>
> *cluster sampling:* Used in large scale descriptive studies involving target populations with geographically dispersed sampling units. The cluster, or primary sampling unit, might represent a hospital or a block within a neighborhood. The elementary sampling unit on which measurements are made might be the patients in the hospital or the residents of the block. Commonly used in epidemiological studies.
>
> *convenience sampling:* Subjects are selected because they easy to identify and happen to be available for participation in the study at a certain time.

purposive sampling: A sample in which the sampling units are deliberately (non-randomly) selected according to certain criteria that are known to be important and are considered to be representative of the population.

quota sampling: Similar to a convenience sample, but the use of controls prevents overloading the sample with subjects having certain characteristics. The controls are established by determining the distribution of the sampling units according to those variables deemed to be important.

random (probability) sampling: A method whereby each sampling unit in the target population has the same, known probability (greater than zero) of being selected in the sample. Neither the sampler nor the sampling unit has any conscious influence over the inclusion in the sample; selection is by chance.

sequential sampling: The sampling units are taken into the study sequentially and the number to be included in the study is not fixed in advance.

stratified random sampling: The target population is subdivided into homogeneous subpopulations. Then, a random sample or a systematic sample is selected from each subpopulation.

systematic sampling: After a random start, every *n*th unit (name on a list, patient in a bed, house on a block) is selected in the order in which the units are arranged.

Sampling unit: The individual members of a target population (humans, animals, plants, inanimate objects etc.).

Science: Organized curiosity.

Significance level: The level of probability at which the null hypothesis will be rejected. A result is statistically significant (by definition) if the null hypothesis is rejected. That is, the probability of the observed result (the *p* value) is below an arbitrary threshold (conventionally set 0.05). The significance level is the threshold value for *p*. Remember that a small *p* value may be the result of a large sample size, so that even small differences or slight correlations are detected. Thus, the observed results may be statistically significant but clinically unimportant.

Statistic: A numerical characteristic of a sample, such as the sample mean and standard deviation.

Theory: A general statement of apparent relationships or underlying principles of certain observed phenomena, which has been verified to some degree by testing specific hypotheses.

Type I error: The error of ejecting the null hypothesis when it is true. For example, if the significance level of a test is set at $\alpha = 0.05$, and the *p* value for a statistical test turns out to be $p = 0.01$, then reject the null hypothesis. The probability of this rejection being in error is 0.01.

Type II error: The error of accepting the null hypothesis when it is false. For example, if the significance level of a test is set at $\alpha = 0.05$, and the *p* value for a statistical test turns out to be $p = 0.25$, then accept the null hypothesis. To calculate the probability of being in error, β, you need the power of the test (often calculated by statistics programs along with *p*). For example, if the power was 0.80, then $\beta = 1 - \text{power} = 1 - 0.8 = 0.20$.

Validity: A criterion for evaluating the quality of a method for measuring variables (often applied to humanistic measurements such as patient satisfaction; surveys and psychological tests are often referred to as measuring "instruments").

construct validity: The extent to which an instrument is consistent with and reflects the theory underlying it.

Content validity: The extent to which an instrument adequately encompasses the pertinent range of subject matter.

face validity: The extent to which an instrument appears as a logical measure of what it is supposed to measure.

predictive validity: The extent to which an instrument is correlated with an objective criterion measure (e.g., job satisfaction with job turnover). This is the most important assessment of validity.

Visual analog scale: A measurement technique designed to obtain data; requires respondents to select a point on a linear scale to indicate the intensity of their feelings or opinions.

Appendix B. Peer Review Checklists

This appendix shows checklists used by reviewers for Respiratory Care journal.

Original Study Checklist

No	?	NA	Introduction
			1. Is the background information adequate to introduce the research problem?
			2. Are the references adequate?
			3. Are specific study objectives or hypotheses stated?
			4. Is the writing in this section clear and concise?
No	**?**	**NA**	**Methods**
			5. Are there outcome variables described for each study objective or hypothesis?
			6. Are there appropriate descriptions of how calculated values were determined?
			7. Are outcome variables that do not relate to the objectives or hypotheses avoided?
			8. Are the measurement procedures appropriate for the study objectives or hypotheses?
			9. Is there enough detail to judge validity and for readers to replicate study?
			10. Is there adequate description of calculated values?
			11. Were appropriate statistical methods chosen for this study design?
			12. Are the tables, figures, and captions adequate?
			13. Is the writing in this section clear and concise?
No	**?**	**NA**	**Results**
			14. Are there complete data for each procedure or test described in the Methods section?
			15. Do the data, or descriptions of the data, appear to be valid?
			16. Are data that do not relate to the study objectives or hypotheses avoided?
			17. Are the tables, figures, and captions adequate?
			18. Does the text avoid presenting the same data as the tables or illustrations?
			19. Is the writing in this section clear and concise?
No	**?**	**NA**	**Discussion and Conclusion**
			20. Is there an explanation of how the results address the problem statement or hypotheses?
			21. Are theoretical and practical aspects of the results discussed?
			22. Do you agree with the interpretation of the results?
			23. Is there a comparison of this study with previously published studies?
			24. Are the references adequate?
			25. Is there a discussion of the limitations of the study?
			26. Is there a section that clearly states the author's conclusions?
			27. Is the writing in this section clear and concise?
No	**?**	**NA**	**Miscellaneous**
			28. Does the paper's title reflect the paper's content?
			29. Is the abstract informative: briefly outlining hypotheses, methods, results, and conclusions?
			30. Is there a Product Sources listing?

Device/Method Evaluation Checklist

Yes	No	?	NA	Introduction
				1. Is the background information adequate to introduce device/method?
				2. Are the references adequate?
				3. Is it stated whether this is a descriptive study, accuracy study, or agreement study?
				4. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Device/Method Description**
				5. Is there enough detail to explain device/method if unfamiliar to readers?
				6. If device: is it explained that it is a new product, commercially available, or prototype?
Yes	**No**	**?**	**NA**	**Evaluation Methods**
				7. Are there appropriate descriptions of how calculated values were determined?
				8. Are the measurement procedures appropriate for the study objectives or hypotheses?
				9. Is there enough detail to judge validity and for readers to replicate study?
				10. Is there adequate description of calculated values?
				11. Were appropriate statistical methods used? See Resp Care 1996;41:1092-1099.
				12. Are the tables, figures, and captions adequate?
				13. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Evaluation Results**
				14. Are there complete data for each procedure or test described in the Methods section?
				15. Are there cost data if appropriate?
				16. Do the data, or descriptions of the data, appear to be valid?
				17. Are data that do not relate to the study objectives or hypotheses avoided?
				18. Are the tables, figures, and captions adequate?
				19. Does the text avoid presenting the same data as the tables or illustrations?
				20. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Discussion and Conclusion**
				21. Is there an explanation of how the results address the study purpose?
				22. Are theoretical and practical aspects of the results discussed?
				23. Do you agree with the interpretation of the results?
				24. Is there a comparison of this study with previously published studies?
				25. Are the references adequate?
				26. Is there a discussion of the limitations of the study?
				27. Is there a section that clearly states the author's conclusions?
				28. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Miscellaneous**
				29. Does the paper's title reflect the paper's content?
				30. Is the abstract informative: briefly outlining hypotheses, methods, results, and conclusions?
				31. Is there a Product Sources listing?

Case Study Checklist

Yes	No	?	NA	Introduction
				1. Is the background information adequate to introduce the case study?
				2. Does the diagnosis satisfy accepted criteria?
				3. Are data supplied to confirm the diagnosis?
				4. Does the report provide an adequate description of the patient's presentation and condition?
				5. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Case Summary**
				6. Is the case summary complete?
				7. Does the treatment make good theoretical and physiological sense?
				8. Is the treatment safe?
				9. If the treatment is new, is it needed (ie, are current methods inadequate)?
				10. If the treatment is new, is it an improvement?
				11. Is the treatment cost effective?
				12. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Discussion**
				13. Is the significance of the case clear?
				14. Is the case uncommon or of exceptional teaching value?
				15. Is the writing in this section clear and concise?
Yes	**No**	**?**	**NA**	**Illustrations**
				16. Do the illustrations add value to the case report?
				17. Are the illustrations clear and well labeled?
Yes	**No**	**?**	**NA**	**Miscellaneous**
				18. Does the paper's title reflect the paper's content?
				19. Is the abstract informative: briefly outlining introduction, case summary, and discussion?

Appendix C. Model Paper

What follows is an actual manuscript that was published in a peer reviewed medical journal (Respiratory Care 2001;46(5):466-474). It is presented in its final draft form, after being reviewed and revised, just before publication. At the next Appendix, you will find a checklist responding to all of the comments the reviewers made. This example will give you a good idea of how to respond to critical review.

ELECTRICAL STIMULATION FOR SWALLOW DISORDERS CAUSED BY CEREBROVASCULAR ACCIDENTS

2/23/2004

AUTHORS AND TITLES

Marcy Freed, M.A. SLP	Speech-Language Pathologist, University Hospitals of Cleveland (formerly Hillcrest Hospital/ Cleveland Clinic Health Systems)
Leonard Freed, Ph.D.	Department of Zoology, University of Hawaii
Robert L. Chatburn, RRT, FAARC	Director, Department of Respiratory Care, University Hospitals of Cleveland
Michael Christian, M.D.	Department of Radiology, Hillcrest Hospital

This is to inform you that certain members of the research team have applied for and received a patent on this technique and device with further claims now pending. As of this date, there has been no money promised or received from any business group. The study was funded in total by the authors and the research team.

Corresponding Author : Ms. Marcy Freed, M.A., SLP
c/o Respiratory Care Department
University Hospitals of Cleveland
11100 Euclid Avenue
Cleveland, OH 44106
216 / 844-7283

Key Words: swallowing, dysphagia, electrical stimulation, cerebrovascular accident, modified barium swallow

Abstract

INTRODUCTION: An estimated fifteen million adults in the United States are affected by dysphagia (difficulty swallowing). Severe dysphagia predisposes to medical complications such as aspiration pneumonia, bronchospasm, dehydration, malnutrition, and asphyxia. These can cause death or increased health care costs from increased severity of illness and prolonged length of stay. Existing modalities of treating dysphagia are generally ineffective and at best it may take weeks to months to show improvement. One common conventional therapy, consisting of the application of cold stimulus at the base of the anterior faucial arch, has been reported to be somewhat effective. We describe an alternative treatment consisting of transcutaneous electrical stimulation applied through electrodes placed on the neck. **OBJECTIVE:** The purpose of this study was to compare the effectiveness of electrical stimulation (ES) treatment to thermal-tactile stimulation (TS) treatment in patients with dysphagia caused by cerebrovascular accident (CVA), and to assess the safety of the technique. **METHODS:** In this controlled study, patients with CVA and possible swallow disorder were alternately assigned to one of the two treatment groups (TS, ES). Entry criteria included a primary diagnosis of cerebral vascular accident (CVA) and confirmation of swallow disorder by modified barium swallow (MBS). TS consisted of touching the base of the anterior faucial arch with a metal probe immersed in ice. ES was administered with modified hand-held battery powered electrical stimulator connected to a pair of electrodes positioned on neck. Daily treatments of each type lasted one hour. Swallow function before and after the treatment regimen was scored from 0 (aspirates own saliva) to 6 (normal swallow) based on substances the patients could swallow during a modified barium swallow. Demographic data were compared with the t-test and Fisher Exact Test. Swallow scores were compared with the Mann-Whitney U test and Wilcoxon Signed Rank Test. **RESULTS:** Both treatment groups

had similar age and gender ($p > 0.27$), correlated co-morbid conditions ($p= 0.0044$), and similar initial swallow score ($p = 0.74$). Both treatment groups showed improvement in swallow score but the final swallow scores were higher in the ES group ($p < 0.0001$). In addition, 98% of ES patients showed some improvement while 27% of TS patients remained at initial swallow score and 11% got worse. These results are based on similar numbers of treatments (average of 5.5 for ES and 6.0 for TS, $p = 0.36$). **CONCLUSIONS:** Electrical stimulation appears to be a safe and effective treatment for dysphagia due to CVA and results in better swallow function than conventional treatment consisting of thermal-tactile stimulation.

Introduction

An estimated fifteen million adults in the United States[1] are affected by difficulty in swallowing (dysphagia). The prevalence of dysphagia in certain diseases may approach 90% (e.g., amyotrophic lateral sclerosis, Parkinson's disease, and certain types of cerebrovascular accidents).[2] Severe dysphagia predisposes to medical complications such as aspiration pneumonia, bronchospasm, dehydration, malnutrition, and asphyxia. These can cause death or increased health care costs from increased severity of illness, prolonged length of stay, readmissions, respiratory support, tracheotomies, and percutaneous enterostomal gastric (PEG) tube placement, with nutritional supplements and equipment[2,3,4] Aside from the physical complications of aspiration, patients often suffer severe depression due to the loss of the swallow function and the disruption of normal activities of daily living.

Existing treatments for dysphagia are unable to restore complete swallow function in patients with the most severe disorders. Physical maneuvers to compensate for the deficiency (such as tucking the chin and suck swallow) are considered generally ineffective.[5,6] Thermal-tactile stimulation, that is, application of cold on the anterior faucial arch[7,8] and biofeedback[9] have success rates that vary from 0 to 83%.[5,9,10,11] Studies reporting high success rates with stroke patients generally do not include the most severe forms of dysphagia in which patients initially aspirate everything, including their own saliva. Often, these studies simply state that improvement was resumption of oral intake, but they do not describe the consistency of the oral intake. The type of oral intake is important because it affects not only hydration and nutrition but also the psychosocial impact on the patient. The minimal goal of treatment should be to achieve

sufficient oral intake to prevent or remove a PEG tube, with its attendant difficulties of reflux aspiration and complications associated with infections. The ultimate goal should be restoration of a normal swallow.

Current modalities have long treatment times: 2-52 weeks (average 15 weeks) for severe dysphagia using tactile and thermal-tactile stimulation[5] and 3-29 weeks using biofeedback.[9] A four-fold increase in pneumonia has been documented during treatment as compared to the post-treatment periods.[5] Lengthy treatment of swallowing disorders is thus risky and may potentially interfere with treatment of other medical problems.

Spontaneous improvement in swallowing may occur in certain acute diseases that cause mild dysphagia.[12] However, in the US only 2% of patients with neurologic disorders and PEG's returned to full oral feed after one year, suggesting that spontaneous improvement is rare for cases of severe dysphagia.[3]

Electrical stimulation has been reported as a treatment for dysphagia.[13,14] Park et al[15] applied electricity through a prosthetic device on the soft palate, aiming to re-educate neural pathways associated with the swallowing reflex. They reported a 50% success rate in improving the swallow of patients already capable of oral feeding. Transcutaneous application of electrical current to the neck with a nerve stimulator has also been successful in improving swallow function but has rarely been used because of assumed concerns for safety[6,16]

We report a new treatment for dysphagia consisting of transcutaneous electrical stimulation applied through electrodes placed on the neck. The purpose of this study was to compare the effectiveness of electrical stimulation (ES) treatment to thermal-tactile stimulation (TS) in patients with dysphagia caused by cerebrovascular accident, and to assess the safety of the technique. Because ES is a more direct stimulus than TS to nerves and muscles associated with swallowing, we hypothesized that ES would result in higher swallow function than TS in patients with comparable conditions of dysphagia. We also monitored patients after treatment to investigate the long-term effects of treatment and the potential for spontaneous recovery.

Methods

The study was conducted at Hillcrest Hospital, a 280 bed acute care hospital in a suburb of Cleveland, Ohio. All new referrals who met entry criteria and signed the consent form were enrolled during the study period. The study period was from 9/23/93-1/24/95. The study population included both inpatients and outpatients. Entry criteria included:

- Primary diagnosis of cerebral vascular accident (CVA).
- Confirmation of swallow disorder by modified barium swallow (MBS).
- Exclusion criteria were:
- Inability to complete at least two consecutive days of therapy.
- Any behavioral disorder that interfered with administration of therapy.
- Significant reflux from feeding tube.
- Dysphagia from drug toxicity.
- Duration of swallow dysfunction did not limit eligibility. Informed consent, as approved by the Institutional Review Board, was obtained for all patients.

Patients with CVA and possible swallow disorder were alternately assigned to one of the two treatment groups (TS, ES) independent of any other information and before being seen by the

speech-language pathologist. After assignment, the speech-language pathologist performed the MBS with a radiologist to determine the severity of the swallow disorder and assign a swallow score (see Assessment Protocol below). Once it was confirmed that the patient did not meet any exit criteria, the treatment regimen was begun. No patients were excluded from the study based on the severity of dysphagia. The Institutional Review Board approved the study and written informed consent was obtained. After the course of treatment was stopped, another MBS was performed and a final swallow score assessed.

Assessment Protocol

The swallow function of all patients was evaluated by a standardized Modified Barium Swallow (MBS)[7,17] with the addition of following the bolus into the stomach to identify potential esophageal reflux that could result in aspiration. Patients were asked to swallow various consistencies of food mixed with barium powder while being observed under fluoroscopy. Food consistencies progressed from thick to thin until aspiration occurred. Penetration was defined as entry of the bolus into the laryngeal vestibule. Aspiration was defined as passage of barium below the level of the vocal cords. The results of the MBS were interpreted as a swallow score according to the criteria listed in Table 1. The swallow score was assigned as follows: The speech therapist would perform the MBS and send the videotape of the procedure to a designated radiologist. The radiologist would then provide an narrative interpretation of the tape in terms of what type of liquid could be safely swallowed. That narrative report was sent back to the speech therapist who then assigned the corresponding score as shown in Table 1. There were three radiologists who assigned scores and they never knew which treatment the patient had at the time of scoring.

The MBS procedure we used was standard except for two items. First, instead of barium paste, we used barium powder because it has less effect on the consistency and taste of the liquid it is mixed with. The idea is to create mixtures of different, realistic consistencies but with as much of the original taste as possible. Paste has a greater tendency to thicken the mixture than powder and also has a more objectionable taste. The second difference was in the order of consistencies presented to the patient. Standard references suggest using thin liquid (eg, water), then pudding, and then cookie.[18] The problem with this order is that thin liquids may be (but are not always) the most easily aspirated. Thus, if the patient aspirates early in the procedure because thin liquid was used first, then (a) the airway becomes contaminated with barium, making visualization of aspiration for other substances difficult and (b) because of the aspiration, the procedure may be terminated without determining what consistency can be safely swallowed.

During the procedure, the speech-language pathologist auscultated the right mainstem bronchus during inspiration. A normal swallow was a single or polysyllabic sound of 1-2 second duration, representing the movement of food through the pharyngeal area into the esophagus, and consisted of only clear breath sounds.[19] This technique enabled the therapist to identify abnormal swallowing or so-called silent aspiration by airway sounds, including rales and rhonchi, during post swallow inspiration. Silent aspiration is a condition in which food or liquid enters the airway but does not produce any obvious signs of aspiration (ie, there is no cough during or after the swallow).[20] The use of auscultation of the right bronchus during inspiration and following ingestion of the food or liquid bolus aided in hearing changes in lung sounds, and changes in rate of respirations which often trigger concern for silent aspiration and justified the medical

necessity of an MBS. Swallow function (by auscultation) was assessed each day of treatment protocol to check for silent aspiration.

Treatment Protocols

General Treatment Protocol

Inpatient treatment (either ES or TS) began within 24 hours of initial evaluation. Duration was one hour per day of treatment and ten minutes of challenge/assessment. If a patient fatigued, treatment was continued later in the day, as often as necessary, to obtain the full hour. Treatment continued on consecutive days until a swallow function score of at least 5 was achieved or the patient was discharged due to insurance constraints. Those patients discharged before achieving a score of 5 avoided a PEG if they could achieve a score of at least 2 on consistency of liquid.

Outpatients were treated three times per week for one hour per treatment. Treatment continued until they achieved a swallow score of 6 or it was judged that no more progress would be made.

Follow-up on patients was based on medical records (for readmission), or consultation with patient, family, physician, or nursing home therapists, for up to 3 years.

Thermal-tactile Stimulation Treatment Protocol

TS was given in three, twenty minute intervals daily. A speech pathologist (one of the authors, MLF) used the standard methodology for TS including verbal coaching. Thermal-tactile

stimulation was applied with a 00 mirror which had been cooled by immersion in ice or lemon ice. The base of the anterior faucial arch was lightly touched with the mirror back. The mirror was removed and patients were asked to close their mouths and attempt to swallow their saliva (dry swallow). TS and verbal coaching continued. If a dry swallow was elicited, the patient was challenged with thickened liquids (pudding viscosity).

Electrical Stimulation Treatment Protocol

ES was administered by a physical therapist in conjunction with a speech pathologist (MLF), using a modified hand-held battery powered electrical stimulator (Staodyn EMS +2, Staodyne Inc., Longmont, CO). Electrodes were placed on the neck in one of two positions (Figure 1) and were repositioned until muscle fasciculations occurred or the strongest contraction was observed during the swallow response. Neuromuscular electrical stimulation consisted of a symmetric rectangular AC current passing between positive and negative snap skin electrodes. Frequency and pulse width were fixed at 80 Hz and 300 microseconds. Current intensity was set to the patient's tolerance and comfort level. Tolerance and comfort differed among individuals. The sensation most people experienced was first a very slight tingling or crawling sensation. As the intensity was increased (in 2.5 mA increments from a start of 2.5 mA up to a maximum of 25.0 mA.) the individuals perceived a strong vibration or the sensation that the electrodes were coming lose from the neck. Most individuals accommodated rapidly enough to the sensations that the intensity could be continuously increased until contractions were consistently audible (designated the therapy current level). When ES was successful in obtaining a voluntary swallow response, the patient was asked to attempt a swallow with a specific oral consistency. Electrical stimulation was delivered at the therapy current for a total of 60 minutes per treatment in the continuous mode, with a 1.0 second pause between each minute.

All patients were monitored continuously for ECG and SpO_2. A drop in SpO_2 of more than 2% was considered a desaturation due to aspiration. Laryngospasm was defined as a spasmodic closure of the glottis with severely limited ability to ventilate. Laryngospasm was judged by the speech therapist, during treatment, based on audible or visible signs of respiratory distress. All recordings were reviewed and interpreted by the Medical Chief of Staff of the acute care facility.

Data Analysis

Unpaired t-tests were used to compare the mean ages and the total number of treatments in the two groups. A Fisher Exact test was used to compare the proportions of females to males in each group. The similarity of co-morbid conditions was evaluated with the Kendall Tau test (ie, if a high proportion of TS patients have a co-morbid condition, do a high proportion of ES patients also have the co-morbid condition and vice versa). The proportions of confounding factors (ie, brainstem vs hemispheric vs multiple strokes) in the two groups were compared with the Chi-Square test. The Mann-Whitney U test was used to compare the initial swallow scores (ie, to determine if the initial degree of dysphagia upon entering the study was the same for both groups) and the distributions of final swallow scores (ie, it to determine if greater improvement was shown by one treatment group). The change in swallow scores (ie, initial versus final) was evaluated with the Wilcoxon Signed Rank Test. Analyses were performed with StatView statistical software with significance set at $p < 0.05$.

Results

One hundred twenty five patients were screened for possible inclusion in the study. Fifteen refused to sign consent after meeting entry criteria, leaving 110 who were enrolled. Table 2 shows that 99 patients completed the study. All TS patients were inpatients. All but six ES patients were inpatients and one was both an inpatient and outpatient. Eleven patients dropped out of the study; 6 had drug toxicity from other treatments, 2 were transferred to other hospitals, and 3 dropped for unrecorded reasons.

The two treatment groups were comparable in terms of mean age and gender distribution and in co-morbid conditions that would affect treatment outcome (Table 2). The condition that would most negatively affect the conventional treatment group was dementia, and the prevalence was identical in both groups. The presence of confounding factors related to the type of lesion (ie, brainstem vs hemispheric stroke vs multiple strokes) was similar in both groups (Table 3). The TS and ES treatment groups had similar distributions of initial swallow score ($p = 0.74$). There were aphasic patients in both groups, but aphasia did not affect their treatment. There were no patients in the study with apraxia of swallowing. There were seven ES vs six TS patients with dysarthria but in no case did dysarthria appear to affect outcome.

Both treatment groups showed improvement in swallow score (Table 4). However, Figure 2 shows that electrical stimulation resulted in more people having higher final swallow scores than thermal-tactile stimulation ($p < 0.0001$). In addition, all but one of ES patients showed some

improvement (98%; the one patient remained at a swallow score of 2) while 17 (27%) of TS patients remained at initial swallow score and four (11%) got worse.

ES Patients with +6 changes progressed from swallow function 0 (completely dysphagic) to swallow function 6 (normal swallow). ES Patients with +5 changes included three that progressed from swallow function 1 (tolerates saliva only) to 6, and six patients that progressed from swallow function 0 to 5. Other step changes, less than +5, include some ES patients that achieved swallow functions 5 or 6, but these patients started with swallow function greater than 1. No TS patient, regardless of initial swallow function, achieved a final swallow function greater than 4. These results are based on similar numbers of treatments (average of 5.5 for ES and 6.0 for TS, $p = 0.36$).

Several focused comparisons illustrate further differences between ES and TS. For patients starting in swallow score 0 and 1, achievement of swallow score to level 2 or higher indicates successful treatment, in that PEG is not required. Only 52% (15 of 29) of TS patients but 95% (41 of 43) ES patients experienced successful treatment ($p < 0.0001$). ES treatments were also more successful than TS treatments based on achievement of complete swallow score 6 (35% of ES patients vs 0% of TS patients, $p < 0.0002$), each starting in swallow score 0 or 1. In addition, four TS patients (11%) required a PEG during treatment. None of the 58 ES patients required a PEG during treatment, and a swallow score of 2 was achieved within 1-2 treatments in all ES patients.

Twenty-five bedside evaluations performed by the therapist (ie, auscultation of the right bronchial tree for evidence of rhonchi or change in ventilatory pattern) were compared with corresponding MBS studies interpreted by a radiologist. Out of the twenty-five comparisons, only one disagreed: the therapist judged silent aspiration that was not confirmed on MBS. This yields the decision matrix shown in Table 6.16

The positive predictive value was 24/25 = 96%; the true positive rate was 24/24 = 100%; the false positive rate was 1/25 = 4%.

Follow-up data show that treatments administered during the study generally persisted (Table 5). Most patients retained their final swallow function for over two years (89 % for ES and 67 % for TS). Loss of swallow function during the post-treatment period for ES patients was based on a new episode of the problem causing dysphagia. None of the TS patients showed improved swallow function, while 4 (14%) of ES patients improved (3 confirmed by MBS). There was a high rate of aspiration (24%) in TS patients in comparison with no aspiration in ES patients. Two of the aspirating TS patients received a PEG.

A total of 318 applications of ES were administered to patients during this study. Not a single case of laryngospasm or decrease in saturation (by pulse oximetry) was observed. No change in heart rhythm occurred based on EKG rhythm strip recordings.

Discussion

The demographic similarities between the two groups (Table 2) indicate that the desired properties of randomization from the same underlying population were in fact achieved for the two treatment groups, despite the fact that a strict randomization scheme was not used.[21] There was, however, one general difference between the two groups: The electrical stimulation group was treated much longer after the stroke than the thermal stimulation group. This is because most of them had already failed conventional therapy, which was the reason they were referred for the study in the first place. The longer the period after the stroke, the less success is expected with dysphagia treatment. Despite this potential bias against the electrical stimulation treatment, that group showed better results than the conventional group.

Bedside evaluations are important in determining the safety of treatment, to estimate the patient's progress during the treatment period, and to justify further MBS studies. In our study, auscultation was used to detect silent aspiration during treatment. The ability to detect aspiration by this method was evaluated by comparison with radiographic evidence of aspiration. However, MBS procedures were only done for patients who were suspected of aspiration (silent or not). Therefore, we collected no data from which negative predictive value could be estimated Yet the high positive predictive value suggests that auscultation deserves further study as a potentially useful screening test for silent aspiration. More research should be done to identify the optimum bedside evaluation technique and to compare its accuracy with the "gold standard", MBS.

Application of electrical stimulation to muscles associated with swallowing links swallowing therapy with physical therapy. A fundamental principle of physical therapy is that disuse of a striated muscle leads to atrophy of that muscle, even if the medical condition leading to disuse has no direct effect on the muscle or associated nerves.[22] Loss of muscle tone is identified by physical therapists as lesser or no measurable contractility or strength. When attempts at exercise alone fail to result in contraction of an atrophied muscle, electrical stimulation may enhance tone, to the point where exercise may strengthen or activate the muscle.

There may be an analogy with dysphagia. A medical condition such as cerebrovascular accident may block the primary neural pathway for swallowing. There are fewer myofibrils per motor unit of the laryngeal muscles relative to larger muscles (4-6 vs 4000), and there are numerous small muscles of this type that participate in the oropharyngeal phase of swallow.[23] In addition, the motor units within each laryngeal muscle tend to fire asynchronously during a normal swallow, contrasting with the more synchronous firing of larger muscles designed for strength. Under this model, even a few days without the typical 600-2,400 normal swallows per day [31,32] could lead to long-term dysphagia. While this design of small muscles might make them more susceptible to failure from lack of use, it is possible that this design can respond more fully to electrical stimulation. Perhaps this is a reason why ES of the neck restores an effective swallow with fewer treatments compared to restoration of appropriate function by ES of other muscles of the body.[24] Alternatively, fewer treatments might be associated with stimulating a reflex since swallowing is a complex action which is usually initiated voluntarily but is always completed as a reflex involving afferent and efferent cranial nerves [29,30] and primary and secondary swallow centers discovered in the cortex.[25] These muscle tone and reflex hypotheses also pertain to

success of ES in treating urinary incontinence.[26] Much research is required to determine whether ES, applied at a sensory level in our study, works via a peripheral nerve, a direct effect on the small muscles, the central nervous system, or a combination of these factors.

Our data directly address issues about safety. ES of the head and neck, portrayed in Darwin[27] for study of expression of emotions, has been the subject of major recent debate about safety.[28] Possible risks include arrhythmia, hypotension, interference with pacemakers, laryngospasm, glottic closure, burns, and tumor growth.[29] However, one successful study that applied external ES of a nerve of the neck had no complications.[16] Other studies also indicate a lack of change in vital signs, EKG, or other adverse effects in patients who received implantable recurrent laryngeal and vagal nerve stimulators used to treat spastic dysphonia and to control epilepsy.[30]

External application of ES with a muscle stimulator within the parameters utilized in our study appears safe, at the sensory level of application. Standard electrode placement in our study purposely avoids the carotid body. In addition, the voltage and current used in our device are lower than used in a standard neuromuscular stimulator assumed by the concerns of other authors for safety.

The most important theoretical risk to ES is laryngospasm. In an animal study, laryngospasm was achieved with repetitive suprathreshold ES, but not with single shock excitation of the superior laryngeal nerve. As stimulus frequency went above 32-64 Hz, there was a decrease in adductor after-discharge and glottic pressure. In our study, suprathreshold levels of stimulation

of the superior laryngeal nerve did not occur because of the level of therapeutic current, limits on the maximum current of the stimulator, and attenuation by soft tissues of the neck. The high frequency stimulation of clinical ES for dysphagia exceeded 64 Hz and may be one of the factors protecting against laryngospasm. In addition, the constant current stimulator automatically dropped the voltage to maintain a constant current dose in the event of decreased electrode or tissue resistance. With these safeguards, a device as configured for our study is apparently safe. The hypothetical concerns about safety are not supported by our data.

Although there are reports in the literature that CVA patients can recover their swallow spontaneously, tube feedings were needed for 15 to 60 weeks Howard et al[3] indicated that 30% of all patients continued on total tube feeding at one year post CVA. The patients who received ES in our study began eating following three treatments and did not require tube feeding thereafter. ES may begin muscle reeducation prior to the beginning of spontaneous recovery and prevent the need for tube feeding.

In an age when extensive efforts are made to reduce costs of health care, the ES protocol can contribute significantly to those efforts. Between 300,000 and 600,000 new cases of dysphagia occur each year in stroke patients.[31] In 1992, the cost of U.S. enteral nutrition in neurologic disease alone exceeded 330 million dollars per year.[3] Since the ES protocol restored a swallow function to a score of 2 within 1-2 days of treatment, a patient admitted to a hospital with CVA, who lost their swallow in association with the underlying medical problem, could eat on their own or with reduced assistance as an in-patient. A few treatment days (1 hr each day) as an in- or out-patient for a total of 6 days would be expected to restore normal swallow in 35% of the

most severe cases of CVA and 45% of all CVA cases. The medical implications of this are reduced amounts of therapy (fewer sessions, less travelling), efficient use of swallowing therapists on the hospital staff or in a nursing home, avoidance of surgery for PEG and attendant complications, avoidance of specialized dietary substances, normal liquid intake, and reduced risk of aspiration pneumonia.

Corrected dysphagia would interfere less with treatments for other medical problems while improving cost effectiveness for health care facilities. The social implication of lower medical bills, and less restricted social activities associated with eating, is a higher quality of life for both the patient and the family.

A potential limitation of this study is while the scoring of swallow function was fairly objective (see Table 1) it does not preclude subjective bias. However, we compared the distribution of final swallow scores of 29 TS patients from our study with that of 53 patients treated with TS by Neumann et al[5]and found no difference (Kolmogorov-Smirnoff KS = 0.2531, p = 0.13). Therefore the difference between TS and ES was not likely caused by bias against TS. The physical evidence of MBS reveals no bias in favor of ES. In addition, the swallow function score we used is no more subjective than the score validated and published by Rosenbeck et al.[32] The major difference with our score is that we do not record the trajectory of the bolus, merely whether it is aspirated or not and the consistency of liquid aspirated. Because consistency affects risk of aspiration, the purpose of the score is to rank the consistency of liquid that can be safely swallowed. This is the type of information referring physicians prefer to see as an interpretation of the MBS procedure because it helps them formulate instructions for the patient.

Conclusions

Transcutaneous electrical stimulation appears to be a safe and effective treatment for dysphagia caused by cerebrovascular accident that results in better improvement in swallow function than thermal-tactile stimulation. Normal swallow function was restored to 35% of the most severely dysphagic patients in less than a week of daily treatment, to 45% of patients at all levels of severity, and the restoration persisted until a new episode of dysphagia occurred. The only limitation of ES is that it cannot be done on patients who talk continuously, such as is found in some severely demented patients. On the other hand, TS treatments require the cooperation of patients in opening their mouths and following verbal commands.

References

[1] Bello J, editor**.** Prevalence of speech, voice, and language disorders in the United States. Communication Facts 1994 Edition. American Speech-Language-Hearing Association:1-4.

[2] Gordon C, Hewer RL, Wade DT. Dysphagia in acute stroke. British Medical Journal 1987;295:411-414.

[3] Howard L, Ament M, Fleming CR, Shike M, Steiger E. Current use and clinical outcome of home parenteral and enteral nutrition therapies in the United States. Gastroenterology 1995;109 (2):355-365.

[4] Hogue CW, Lappas GD, Creswell LL, et al. Swallowing dysfunction after cardiac operations. The Journal of Thoracic and Cardiovascular Surgery 1995;110(2):517-522.

[5] Neumann S, Bartolome G, Buchholz D, Prosiegel M. Swallowing therapy of neurologic patients:correlation of outcome with pretreatment variables and therapeutic methods. Dysphagia 1995;10:1-5.

[6] Langmore S, Miller RM. Behavioral treatment for adults with oropharyngeal dysphagia. Arch Phys Med Rehabil 1994;75:1154-1159.

[7] Logemann J. Evaluation and treatment of swallowing disorders. San Diego: College Hill Press; 1983:134-223.

[8] Lazzarra G, Lazarus C, Logemann JA. Impact of Thermal Stimulation on the Triggering of the Swallowing Reflex. Dysphagia 1986:1:73-77.

[9] Crary M. A direct intervention program for chronic neurogenic dysphagia secondary to brainstem stroke. Dysphagia 1995:10:6-18.

[10] Rosenbeck JC, Robbins J, Fishback B, Levine RL. Effects of thermal stimulation on dysphagia after stroke. Journal of Speech and Hearing Research 1991;34:1257-1268.

[11] Singh V, Brockbank MJ, Frost RA, Tyler S. Multidisciplinary management of dysphagia: the first 100 cases. The Journal of Laryngology and Otology 1995;109:419-424.

[12] Barer DH. The natural history and functional consequences of dysphagia after hemispheric stroke. Journal of Neurology, Neurosurgery and Psychiatry 1989;52:236-241.

[13] Freed M. Christian MO, Beytas EM, Tucker H, Kotton B. Electrical stimulation of the neck: A new effective treatment for dysphagia. Dysphagia 1996;11:159.

[14] Chatburn RL, Freed M. Electrical stimulation for treatment of dysphagia in children failing conventional therapy. Respir Care 2000;45(8):1009.

[15] Park CL, O'Neill PA, Martin DF. A pilot exploratory study of oral electrical stimulation on swallow function following stroke: an innovative technique. Dysphagia 1997;12:161-166.

[16] Larsen GL. Conservative management for incomplete dysphagia paralytica. Arch Physician Medical Rehabilitation 1973;54:180-185.

[17] Ott DJ, Pikna LA. Clinical and Videofluoroscopic Evaluation of Swallowing Disorders. AJR 1993;507-513.

[18] Logemann JA. Evaluation and treatment of swallowing disorders. Austin: Pro-Ed, 1998:177.

[19] Leonard R, Kendall K. eds. Dysphagic Assessment and Treatment Planning. San Diego: Singular Publishing Group; 1997.

[20] Carrau RL, Murray T. Comphrehensive management of swallowing disorders. San Diego, Singular Publishing Group Inc., 1999:72.

[21] Fisher RA. The Design of Experiments. Edinburgh: Oliver & Boyd; 1935:13-29.

[22] Gordon T, Mao J. Muscle Atrophy and Procedures for Training After Spinal Cord Injury. Physical Therapy1994;74(1):50-59.

[23] West JB. Best and Taylor's Physiological Basis of Medical Practice, 12th Edition. Baltimore: Williams & Wilkins; 1991.

[24] McQuain MT, Sinaki M, Shibley LD, Wahner HW, Ilstrup DM. Effect of electrical stimulation on lumbar paraspinal muscles. Spine 1993;18:1787-1792.

[25] Hamdy S, Aziz Q, Rothwell J, et al. The cortical topography of human swallowing musculature in health and disease. Nature Medicine 1996;2 (11):1217-1224.

[26] Fall M, Lindstrom S. Electrical stimulation: a physiologic approach to the treatment of urinary incontinence. Urologic Clinics of North America 1991;18:393-407.

[27] Darwin C. The Expression of the Emotions in Man and Animals,3rd ed. New York: Oxford University Press. 1998:21,278-309.

[28] Huckabee M. The risks of good intentions: neuromuscular electrical stimulation. Swallowing and Swallowing Disorders. Letter to Editor. Journal of the American Hearing and Speech Association, May 1997.

[29] Friedman M, Wernicke JF, Caldarelli DD. Safety and tolerability of the implantable recurrent laryngeal nerve stimulator. Laryngoscope 1994;104:1240-1244.

[30] Suzuki M, Sasaki CT. Laryngeal spasm: A neurophysiologic redefinition. Ann Otol 1977;86:150-157.

[31] Doctor's Guide. http://www.pslgroup.com/dg/CE822.htm

[32] Rosenbeck JC, Robbins JA, Roecker EB, Coyle JL, Wood JL. A penetration-aspiration scale. Dysphagia

Table 1. Swallow function scoring system. This system identifies the consistency of liquid that the patient can swallow without aspiration.

Swallow Function Score	**Safe Liquid Consistency**	**Clinical Implication**	**Level of Swallow Deficit**
0	Nothing safe (aspirates saliva)	No solid or liquid is safe	Profound
1	Saliva	Same as above (candidate for PEG)	Profound
2	Pudding, Paste, Ice Slush		Significant
3	Honey consistency (liquid with thickener or premixed product like Resource brand liquid nourishment)		Moderate
4	Nectar consistency (pureed fruit juice such as apricot, peach, pear)		Mild
5	Thin liquids (eg, cream soups, orange juice, carbonated beverage)	No coffe, tea, thin juice (eg, apple), or water	Minimal
6	Water	All liquids tolerated	Normal

Table 2. Treatment groups with respect to demography and health. Patients often had more than one co-morbid condition (ie, proportions were not mutually exclusive).

	Treatment		
Variable	thermal stimulation (n = 36)	electrical stimulation (n = 63)	p value
Average age	78.1	75.7	0.27
Maximum age	91	101	-
Minimum age	65	49	-
Percent female	44	48	0.83
Co-morbid conditions*			
Multiple CVA	0.08	0.11	
Coronary Artery Disease	0.08	0.08	
Congestive Heart Failure	0.14	0.08	
Chronic Obstructive Pulmonary Disease	0.06	0.05	
Hypertension	0.17	0.19	
Dementia	0.03	0.03	
Diabetes Mellitus	0.06	0.08	
Parkinson's Disease	0.00	0.02	
Cancer	0.25	0.10	
Multiple Sclerosis	0.03	0.00	

* proportions were significantly correlated by Kendal's Tau; p = 0.0044

Table 3. Frequencies of various types of lesions. Not all patients were evaluated for type of lesion. The proportion of observations in the different categories is not significantly different than is expected from random occurrence ($p = 0.183$).

Treatment	Brainstem	Hemispheric	Multiple Strokes
electrical stimulation	7	29	24
thermal-tactile stimulation	1	19	8

Table 4. Mean (SD) swallow scores before and after treatment.

	Initial Swallow Score	**Final Swallow Score**
electrical stimulation	0.76 (1.04)	4.52 (1.69)
thermal-tactile stimulation	0.75 (1.20)	1.39 (1.13)

Table 5. Proportions of patients in post-treatment categories.

Category	Treatment: thermal stimulation (n=33)	Treatment: electrical stimulation (n=52)
No change for > 2 yr., alive	0.061	0.289
No change for < 2 yr., lost *	0.242	0.269
No change for < 2 yr., died *	0.364	0.250
Improved within 2 yr.	0 .000	0.077 †
Aspiration or PEG	0.242	0 .000
New episode of dysphagia ‡	--	0.115
Received ES after TS ‡	0.091	--

* average time of follow-up greater than 1 yr.

† proportion is 0.143 of 28 ES patients with final swallow function less than 6

‡ full swallow function restored after ES

Table 6. Analysis of agreement between bedside assessment of silent aspiration and results of MBS.

	MBS Interpretation	
Bedside Evaluation	Aspiration Present	Aspiration Absent
Aspiration Present	**24**	**1**
Aspiration Absent	**0**	**0**

Figure Legends

Figure 1. Diagram of the throat showing placements for pairs or snap electrodes. One of two placements was used: (A) on either side of the midline, above the lesser horns of the hyoid bone, on the digastric muscle. (B) on either side of the midline (preferably on right side) with upper electrode above lesser horns of the hyoid bone, on the digastric muscle, and lower electrode on the thyrohyoid muscle at the level of the top of the cricothyroid cartilage. Position A was used for patients with tracheostomies or those whose anatomy prevented using the other position. Position B was used for everyone else.

Figure 2 Distributions of initial and final swallow scores for electrical stimulation (ES) and thermal-tactile stimulation (TS) treatment groups. A higher score means better swallow function. Initial swallow scores for the two groups were similar ($p = 0.74$). Both groups showed improvement in score (TS, $p = 0.0048$; ES, $p < 0.0001$). The ES group had higher final scores ($p < 0.0001$).

Acknowledgments: The authors wish to acknowledge the contributions of the following *(alphabetically):*

Marie Asmar P.T.

Erol Beytas M.D.,

Robert E. Botti MD,

Rebecca Cann, PhD,

Kenneth Hawk P.T.,

Bernard Kotton MD,

Nancy Lynam Davis,

Joan Rothenberg, M.D.

Howard Tucker, MD,

Figure 1

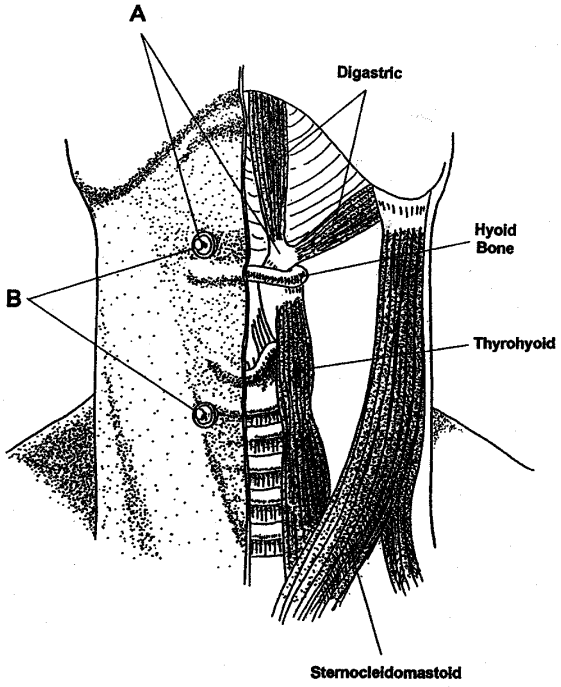
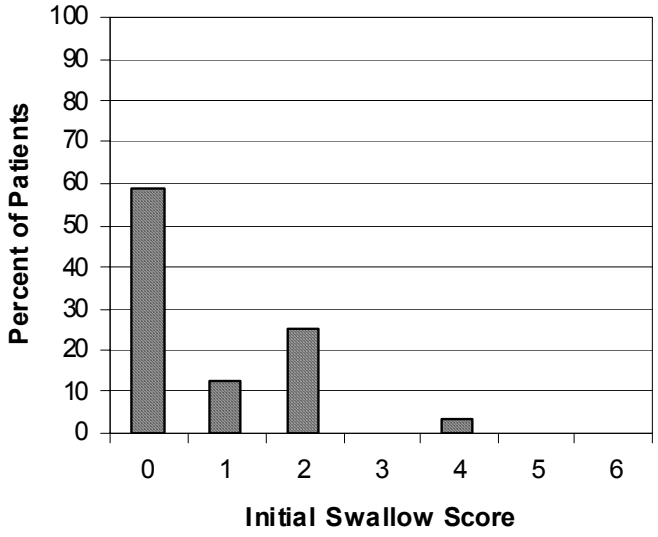
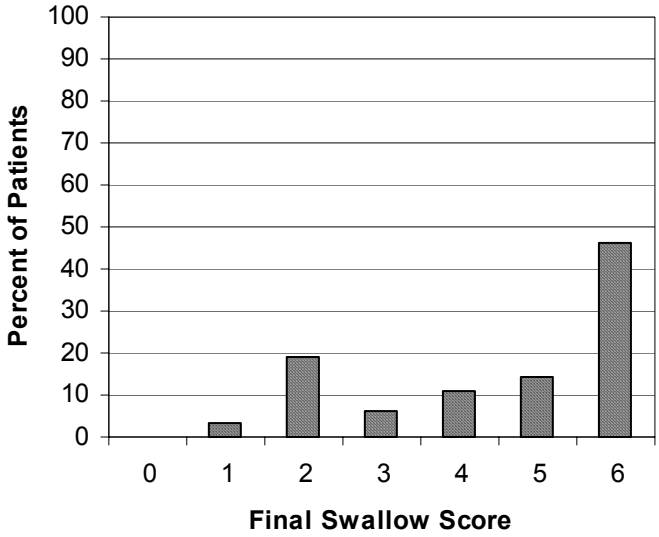


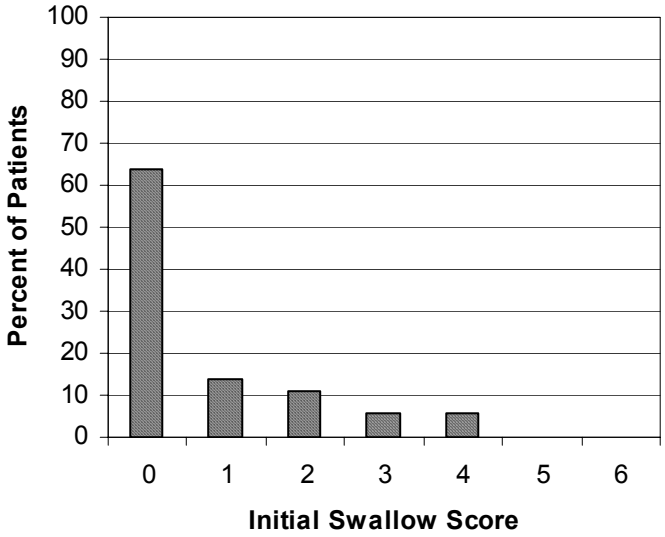
Figure 2

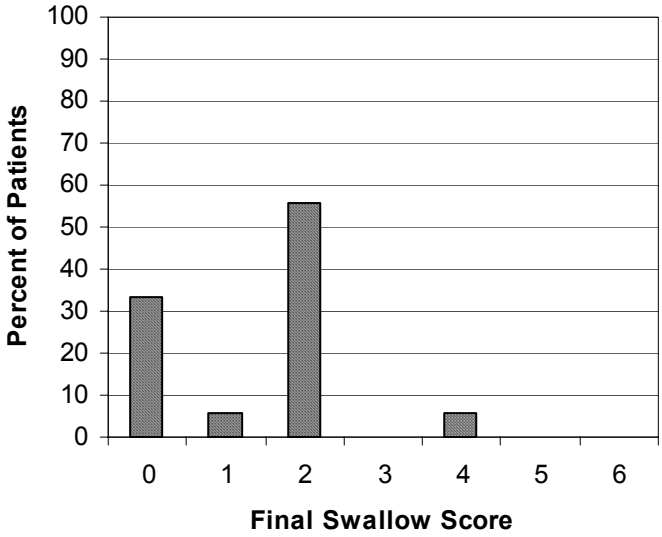
Electrical Stimulation





Thermal-Tactile Stimulation





Appendix D. Response to Reviewers

This appendix is part of an actual response to review of the model paper in the previous Appendix. The reviewers' comments are in italics, my response is below.

1. Abstract

I cannot determine if the authors are talking about dysphagia in general or oral/pharyngeal dysphagia in particular. They need to define "dysphagia".

We believe the distinction between oral and pharyngeal dysphagia is irrelevant to our study, as reflected in this statement from a leading textbook "Pathology that may cause dysphagia includes abnormal movement of the tongue in forming the bolus and initiating deglutition, residual barium that pools in the valleculae or pyriform sinuses, and aspiration of barium into the airway."[†] We will stick with the standard definition of dysphagia that anyone can look up in a medical dictionary. Our treatment outcome is not expected to have been affected by oral vs pharyngeal problems.

A cold stimulus is not applied to the back of the throat, it is applied to the anterior faucial pillar which is not at the back of the throat-the pharyngeal constrictors are at the back of the throat.

The exact position was described in the Methods section. We have changed the Abstract text to match.

Transcutaneous electrical stimulation was applied at the neck.

We have changed the text as suggested.

Thermal stimulation is no longer considered the appropriate term to use, the commonly used term is thermal-tactile stimulation. This needs to be changed throughout.

The text was changed as suggested.

[†] Carrau RL, Murray T. Comphrehensive management of swallowing disorders. San Diego, Singular Publishing Group Inc., 1999:72.

2. Background

I have no idea what the authors mean by the sentence "Human swallowing is compromised by additional modification to the throat to facilitate speech"

This whole section describing the physiology of normal swallow was removed.

The tracheal opening is not back in the throat. Additionally, it is inferior to the pharynx which is likely what the authors mean when they use the term "throat".

Steadman's Concise Medical Dictionary (3rd edition, page 672) defines pharynx as "…the throat, the joint opening of the gullet and windpipe".

The tongue generally goes against the alveolar ridge, not the hard palate.

We have reworded the text to read "Next, the velum elevates, the lips and buccal muscles contract, and the posterior aspect of the tongue depresses. The remainder of the tongue presses against the hard palate as it propels the bolus toward the oropharynx."‡

Elevation of the larynx does not force the food through the fauces, and the fauces are not in the pharynx.

We stated that "Elevation of the larynx ***and backward movement of the tongue*** forces the food through the isthmus of the fauces …" Certainly the tongue movement does generate the force to propel the food bolus. Our text did continue "… in the pharynx." which should have been "into the pharynx."

The correct term is epiglottic inversion or downfolding, not backward bending.

The text has been corrected to "…epiglottic inversion..".

Laryngeal elevation plays a major role in reducing the likelihood of penetration; the authors list only glottic closure and epiglottic inversion which are minimally effective without elevation.

‡ Perlman AL. Schulze-Delrieu K. Deglutition and its disorders. San Diego: Singular Publishing Group, Inc., 1998: 16.

The text has been changed to "Food is kept from entering the nasal cavity by elevation of the soft palate. Food is kept from entering the trachea by elevation and anterior movement of the hyoid and larynx accompanied along with closure of the larynx by the true vocal folds, the false vocal folds, and the epiglottis."

The reflexive stage of the swallow is triggered in the region of the anterior faucial arch, we don't know if it is triggered at the arch.

We have changed the text from "…at the faucial arch.." to "…in the region of the faucial arch…"

material formed during the oral stage not at the oral stage.

The text has been changed as suggested.

On page 5, the bottom line is a completely erroneous statement regarding the cranial nerves. Of anything in the text, this error is the most blatant indication that the authors are ill informed. …no clinician who diagnoses and/or treats dysphagia, let alone one who attempts to be a researcher, should make such a gross mistake.

Our text states that "Cranial nerves V, VII, IX, and XII are afferent and cranial nerves IX and X are efferent."

According to references cited by Logerman[§] "…the sensory portion of the pharyngeal swallow is carried by cranial nerves IX, X, and XI". "Nerves V, VII, and XII have been identified as possible contributors to the afferent portion". "The motor portion is carried by nerves IX and X." Miller[**] classifies nerves V, IX, X as sensory-motor, and VII as sensory.

These references (which we have added) validate our statements. However, this whole section was removed along with the description of normal and abnormal swallow physiology.

[§] Logerman JA. Evaluation and treatment of swallowing disorders. 2nd edition, Austin: Pro-Ed, 1998:31.

[**] Miller AJ. The neuroscientific principles of swallowing and dysphagia. San Diego: Sungular Publishing Group, Inc., 1999: 35-38, 49.

Page 6, the epiglottis does not open.

We did not say the epiglottis opened. The sentence was poorly phrased and easily taken out of context. We have reworded it.

Aspiration is not synonymous with choking…

Sentence reworded.

Page 8, electrical stimulation statement needs a citation.

We have added two references.[††,‡‡]

3. Methods

The modified barium swallow procedure was not "standard".

The procedure we used was as described in our references except for two things. First, instead of barium paste, we used barium powder because it has less effect on the consistency and taste of the liquid it is mixed with. The idea is to create mixtures of different, realistic consistencies but with as much of the original taste as possible. Paste has a greater tendency to thicken the mixture than powder and also has a more objectionable taste.

The second difference was in the order of consistencies presented to the patient. Standard references§§ suggest using thin liquid (eg, water), then pudding, then cookie. The problem with this order is that thin liquids "…may be, but are not always, the most easily aspirated…"[§§]. Thus, if the patient aspirates early in the procedure because thin liquid was used first, then (a) the airway becomes contaminated with barium, making visualization of aspiration for other substances difficult and (b) because of the aspiration, the procedure may be terminated without determining what consistency can be safely swallowed.

[††] Chatburn RL, Freed M. Electrical stimulation for treatment of dysphagia in children failing conventional therapy. Respir Care 2000;45(8):1009.

[‡‡] Freed M. Christian MO, Beytas EM, Tucker H, Kotton B. Electrical stimulation of the neck: A new effective treatment for dysphagia. Dysphagia 1996;11:159.

[§§] Logemann JA. Evaluation and treatment of swallowing disorders. Austin: Pro-Ed, 1998:177.

There was no control for site of lesion, previous strokes, presence of aphasia, apraxia, or dysarthria, concomitant medical condition, time post stroke or patient age and gender.

The reviewer is suggesting that these are confounding factors. Here are our thoughts:

- *site of lesion:* this refers to the location in the brain that the stroke affected (ie, brainstem vs cerebellar vs right or left hemisphere). For example brainstem strokes are more likely to result in dysphagia than hemispheric strokes but the latter are more common. A bias could have occurred if the proportion of brainstem vs hemisphere lesions was significantly different between the electrical stimulation and the thermal stimulation group. Another potential confounding factor is whether the patient had multiple strokes, which makes recovery from dysphagia more difficult. We have reviewed the available data (not all patients had CT or MRI scans) and found the following:

Treatment	Brainstem	Hemisphere	Multiple Strokes
electrical stimulation	7	29	24
thermal stimulation	1	19	8

  The proportion of observations in the different categories is not significantly different than is expected from random occurrence (Chi-square, $p = 0.183$). This analysis strengthens our conclusions and it has been added to the text.

- *aphasia:* there were aphasic patients in both the electrical stimulation and thermal stimulation groups, but aphasia did not affect their treatment
- *apraxia:* there were no patients in this study with apraxia of swallowing
- *dysarthria:* there were 7 e-stim vs 6 thermal stim patients with dysarthria but in no case did dysarthria affect outcome
- *concomitant medical condition:* Table 2 lists comorbid conditions with proportions and shows that both groups were comparable on conditions that would affect treatment outcome. The condition that would most negatively affect the conventional treatment group was dementia, and the prevalence was identical in both groups.
- *time post stroke:* the longer the period after the stroke, the less success is expected with dysphagia treatment. The electrical stimulation group was treated, in general, much longer after the stroke than the thermal stimulation group. This is because most of them had already failed conventional therapy, which was the reason they were referred for the study in the first place. So the bias would have been against the electrical stimulation group, which only strengthens our actual findings.

- *patient age and gender:* age and gender are presented in Table 2 and the statistical comparison shows not difference between groups.

…the clinical professional is called a Speech-Language Pathologist, not a speech therapist.

We have changed the text.

Reflux does not cause aspiration but it can result in aspiration.

The word cause was switched to result.

How was maximal pharyngeal contraction determined?

We have reworded the text to say that intensity was increased until contractions were consistently audible.

I am fascinated by the statement that electrical stimulation was delivered for a total of 60 continuous minutes with a 1 second pause between each minute. How did muscle fatigue play into this equation.

The level of electrical stimulation used increased muscular tone but did not generally cause contractions. Therefore, fatigue was a function of how many times the patient voluntarily swallowed oral intake. Thus, fatigue rarely reached the level that the risk of aspiration caused the treatment to be stopped.

Appendix E. Answers to Questions

Chapter 1

Definitions

- *Basic research:* Seeks new knowledge rather than attempting to solve an immediate problem
- *Applied research:* Seeks to identify relationships among facts to solve an immediate problem
- *JCAHO:* The Joint Commission on Accreditation of Health Care Organizations.
- *Quality assurance:* Delivery of optimal patient care with available resources and consistent with achievable goals.

True or False

1. True
2. True

Multiple Choice

1. a

Chapter 2

Definitions

- *IRB*: Institutional Review Board; panel of experts who evaluate and approve research protocols involving humans.
- *Informed consent:* Informed consent is the voluntary permission given by a person allowing himself to be included in a research study after being informed of the study's purpose, method of treatment, risks, and benefits.

True or False

1. True
2. False

Multiple Choice

1. d
2. f
3. b

Chapter 3

Definitions

- *Disease management:* the systematic, population based approach to identify patients at risk, intervene with specific programs, and measure outcomes.
- *Continuous quality improvement:* a cycle of activities focused on identifying problems or opportunities, creating and implementing plans, and using outcomes analysis to redefine problems and opportunities.
- *Outcomes research:* the scientific study of the results of diverse therapies used for particular diseases, conditions, or illnesses.
- *Evidence-based medicine:* an approach to practice and teaching that integrates pathophysiological rationale, caregiver experience, and patient preferences with valid and current clinical research evidence.
- *Benchmarking:* a continuous process of measuring products, services, and practice against one's toughest competitors.

True or False

1. False
2. True
3. False
4. True
5. True

Multiple Choice

1. b
2. c
3. a
4. b
5. d

Chapter 4

Definitions

- *Hypothesis:* a short statement that describes your belief or supposition about a specific aspect of a research problem.
- *Rejection criteria:* a set of criteria set up before the experiment and used to test the hypothesis.

True or False

1. False
2. True

Multiple Choice

1. d
2. e

Chapter 5

Definitions

- *Inductive reasoning:* reasoning from specific observations to general theories.
- *Deductive reasoning:* reasoning from general theories to specific observations.
- *Operational definitions:* terms based on specific operations, observations, or measurements used in the experiment.
- *Feasibility analysis:* judging the overall practicality and worth of a proposed research project.

True or False

1. True
2. True
3. False

Multiple Choice

1. a
2. f
3. c
4. d

Chapter 6

True or False

1. False
2. False
3. True
4. True
5. False

Multiple Choice

1. a
2. c
3. e
4. b
5. d

Chapter 7

Definitions

- *Assessable population:* The collection of cases available to the investigator as defined by the study criteria.
- *Target population:* The entire collection of cases to which research results are intended to be generalized.
- *Sample:* A subset of the population.

- *Variable:* An entity that can take on different values
- *Independent variable:* The manipulated variable; the treatment.
- *Dependent variable:* The measured variable; the outcome of the treatment.
- *Nuisance or confounding variable:* Extraneous (usually uncontrollable) variable that can affect the dependent variable.
- *Placebo:* A treatment designed to appear exactly like a comparison treatment, but which has no active component.
- *Hawthorne effect:* Psychosomatic effects caused by the subject's awareness of being in a study.

True or False

1. False
2. True
3. True
4. False
5. True
6. True
7. False
8. True

Multiple Choice

1. a
2. e
3. b
4. c
5. d

Chapter 8

True or False

1. True
2. False
3. True

4. False
5. True
6. False
7. False

Chapter 9

Definitions

- *Systematic errors:* errors that occur in a predictable manner and cause measurements to consistently under- or over-estimate the true value.
- *Random errors:* errors that occur in an unpredictable manner due to uncontrollable factors.
- *Accuracy:* the maximum difference between a measured value and the true value, often expressed as a percentage of the true value.
- *Precision:* the degree of consistency among repeated measurements of the same variable.
- *Resolution:* the smallest incremental quantity measurable.
- *Calibration:* the process of adjusting the output of a device to match a known input, thus minimizing the systematic error.
- *Calibration verification:* the process of measuring a known value with a calibrated device and making a judgment of whether or not the error is acceptable.

True or False

1. False
2. True
3. False
4. False
5. True
6. True
7. True

Multiple Choice

1. b
2. d
3. b
4. a
5. f
6. b
7. c
8. d
9. e

Chapter 10

Definitions

- *Qualitative variable:* a categorical variable not placed on a number scale.
- *Quantitative Variable:* a variable measured using a scale of numbers.
- *Discrete variable:* a variable with gaps or interruptions such as a variable measured with whole numbers.
- *Confidence interval:* the range of values believed to contain the true parameter value.
- *Error interval:* the range of values expected to contain a given proportion of all future individual measurements at a given confidence level.
- *Null hypothesis:* states that no difference or no association exists.
- *Alternate hypothesis:* states that there is a difference or association.
- *p value:* the probability that the observed results or results more extreme would occur if the null hypothesis were true and the experiment were repeated many times.
- *Significance level(alpha):* The predetermined level of probability at which the null hypothesis will be rejected.
- *Type I error:* The error of rejecting the null hypothesis when it is false.
- *Type II error:* The error of accepting the null hypothesis when it is false.
- *Power:* The probability of rejecting the null hypothesis when it is false.

True or False

1. False
2. True
3. False
4. True
5. True
6. True
7. False
8. True
9. False
10. True

Multiple Choice

1. a
2. d
3. c
4. b
5. a
6. b
7. c
8. b
9. a
10. c
11. a
12. b
13. c
14. d
15. e
16. b
17. a
18. b
19. a

20. c
21. c
22. b
23. d
24. b
25. c
26. a
27. c
28. b
29. devils, demons, death, and damnation

Chapter 11

Definitions

- *Contingency table:* a table used to display counts or frequencies of two or more nominal variables.
- *Proportion:* the number of objects of a particular type divided by the total number of objects in the group.
- *Percentage:* the proportion multiplied by 100%.
- *Ratio:* the number of objects with a given characteristic in a group divided by the number of objects in the group without the characteristic.
- *Odds:* the probability of an event occurring divided by the probability of the event not occurring.
- *Rate:* the quantity of something occurring per unit of time.
- *Sensitivity:* the ability of a diagnostic test to correctly identify patients with the condition of interest
- *Specificity:* the ability of a diagnostic test to correctly identify patients who do not have the condition of interest.
- *Positive predictive ability:* the probability that the condition of interest is present then the diagnostic test is positive
- *Negative predictive ability:* the probability that the condition of interest is absent when the diagnostic test is negative.
- *Receiver operating characteristic curve:* a plot of the true positive rate against the false positive rate for a diagnostic test over a range of possible cut-off values.

True or False

1. True
2. False
3. True

Multiple Choice

1. a
2. c
3. d
4. b

Chapter 12

Multiple Choice

1. a
2. b
3. c
4. d
5. e

Chapter 13

Multiple Choice

1. a
2. d
3. b
4. c
5. a
6. b
7. c
8. c
9. a

10. d
11. b
12. c

Chapter 14

True or False

1. True
2. False
3. True
4. True
5. True
6. False
7. False

Chapter 15

True or False

1. True
2. True
3. False
4. False
5. False

Chapter 16

True or False

1. True
2. True
3. False
4. False

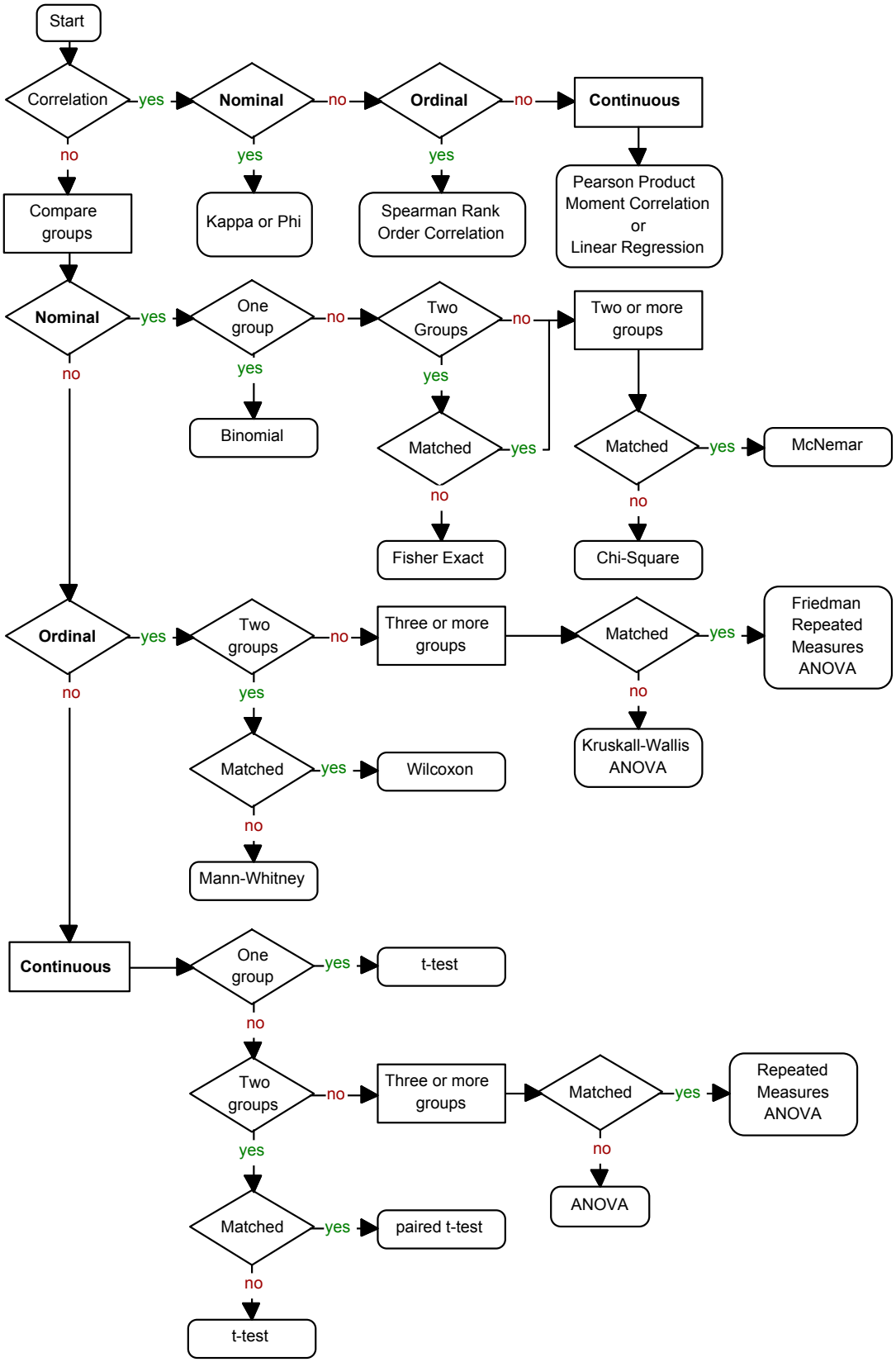
5. True
6. True

Chapter 17

True or False

1. True
2. False
3. True
4. True

Appendix F. Statistics Selector



INDEX

A

accuracy
  definition, 80
  diagnostic, 172
agreement
  strength of for nominal data, 175
agreement interval, 133, 138
  definition, 135
  equation, 135
aliasing, 101
alpha, 145
  definition, 150
anemometer, 94
ANOVA
  Friedman Repeated Measures, 188
  Kruskall-Wallis, 187
  one way, 205
  one way repeated measures, 210
  two way, 206
  two way repeated measures, 212
ANOVA (Analysis of variance), 57
assess quality of life
  example study, 22

B

benchmarking, 23
  indicators, 24
beta
  definition, 150
bias, 82
binomial test, 176

C

Chi-Squared test, 181
clinical trial
  example study, 22
coefficent of determination
  definition, 121
coefficient of variation
  definition, 118
confidence interval
  equation, 130
  table of factors, 131
contingency table, 167, 184
continuous (level of measurement), 107
correlation, 64
  coefficient, 119
  for nominal data, 174
  for ordinal data, 184
  Pearson *r*, 120
  strength of, 120
cost effectiveness, 18
cost identification, 18
cost minimization, 18
cost utility, 18
cost-utility analysis
  example study, 22
CQI (Continuous Quality Improvement), 4
  definition, 15
crossover design, 56

D

damping
  effects on system response, 88
decision analysis, 18
disease nanagement, 15

E

effect size, 153
effectiveness, 20
efficacy, 19
error
  constant, 85
  loading, 88
  operator, 89
  proportional, 85
  random, 80
  range, 85
  systematic, 80
  total, 82
  Type I, 150
  Type II, 150
error interval
  definition, 132
  equation, 133
  plot, 140
ethics

respiratory care, 11
evidence-based medicine
definition, 16
experiment
characteristics of, 53

F

F Ratio test, 193
false negative rate, 171
filter, 100
Fisher Exact test, 178
frequency response, 87, 100

G

gain, 100
Gaussian (curve), 114

H

histogram, 111
hypothesis
definition, 27
research, 145
research, definition, 149
statistical, 149
hypothesis testing, 144
hysteresis, 85

I

imprecision. *See also* precision
inaccuracy. *See also* accuracy
inaccuracy interval, 133, 138
definition, 134
equation, 135
informed consent
background, 8
definition, 8
revocation, 9
interaction, 60
inverse estimation, 143
IRB (Institutional Review Board)
approval, 7
components of proposal, 8
composition, 7
function, 6
protocol outline, 68

J

JCAHO (Joint Commission on Accreditation of Health Care Organizations), 4

K

Kappa, 174
Kolmogorov-Smirnov test, 139, 192

L

likelihood ratio
definition, 172
linearity, 83

M

Mann-Whitney U test, 186
matched data, 160
McNemar's test, 179
mean
definition, 115
median
definition, 115
meta-analysis, 17
mode
definition, 115

N

negative predictive value
definition, 172
noise, 89
nominal (level of measurement), 106
nonlinearity. *See also* linearity
normal (curve), 114
standard normal curve, 125
normality, testing for, 192
null hypothesis, 144

O

ordinal (level of measurement), 107
outcomes research
definition, 16
outliers
treatment of, 140

P

***p* value**
definition, 150
paired data, 160
paired *t* test, 57
parameter, 52, 108
Pearson *r*, 120, 195
peer review, 226
percentage
definition, 167
percentiles plot, 112
Phi, 175
pie chart, 112
placebo, 52
pneumotachometer, 93
point estimates, 130
population
accessible, 49
definition, 105
target, 49
positive predictive value
definition, 172
power
definition, 150
nomogram, 153
power analysis, 152
precision
definition, 82
pressure gauge
Bourdon, 91
diaphragm, 91
piezoelectric, 92
probability distribution, 122
professional conduct. *See* ethics, respiratory rare
proportion
definition, 167

Q

qualitative methods, 19
qualitative research, 16
quality assurance
definition, 4
quality of life, 19
quality-adjusted life years, 19
quantitative methods, 19

R

range, 117
rate
definition, 168
ratio
definition, 167
reasoning
deductive, 34
inductive, 33
regression
logistic, 197
multiple linear, 197
simple linear, 196
reliability
intra-, inter- rater, 174
research
applied, 2
basic, 2
clinical trials, 2
research design
types of, 53
response time, 86
ROC curve (receiver operating characteristic curve), 173
rotameter, 92
rule of threes, 159

S

sample, 49
definition, 105
selecting, 50
sample size
cost control, 159
for confidence intervals, 158
for difference between means (using CV), 157
for difference between means (using S), 155
for difference between proportions, 158
nomogram, 153
rules of thumb, 155
unequal groups, 158
sampling distribution
definition, 127
scattergram, 119
scientific method
definition, 27

sensitivity, 83
definition, 171
significance level, 145
definition, 150
skewness, 114
Spearman Rank Order Correlation coefficient, 184
specificity
definition, 171
spirometer, 95
standard deviation
definition, 117
standard error of the mean
definition, 128
statistic, 52
vs a parameter, 108
statistical significance
vs clinical importance, 160
study designs
types, 29

T

***t* distribution**, 129
***t* statistic**
equation, 129
***t* test**
one sample, 200
paired, 202
unpaired, 200
tolerance interval, 133, 138
definition, 133
equation, 133
tolerence interval
table of factors, 133, 135
true negative rate, 171
true positive rate, 171

U

unpaired *t* test, 57
U-tube manometer, 90

V

validity
external, 61
internal, 61
threats to, 61
variable, 52
continuous, 106
definition, 105
dependent, 52
discrete, 106
independent, 52
nuisance, 52
qualitative, 106
quantitative, 106
variance
definition, 117

W

Wilcoxon Rank Sum test, 186
Wilcoxon Signed Rank test, 187

Z

z score
and standard normal curve, 125
definition, 118
equation, 128
z test, 177

Order Form

Email orders: rlc6@po.cwru.edu

Postal orders:

Mandu Press Ltd
PO Box 18284
Cleveland Heights, OH 44118-0288, USA

Please send _____copies of *Handbook For Healthcare Research* for **$59.95** each to the address below.

Sales Tax: Please add 7% for orders shipped to Ohio addresses.

Shipping by air

US: $4.00 for first book or disk and $2.00 for each additional item.

International: $9.00 for first book or disk and $5.00 for each additional item (estimate).

Payments must accompany order. Allow 3 weeks for delivery.

Name_____

Address_____

_____

City _____

State _____ **Zip**_____