

**DR. SCHAEFER'S**  
**EVIDENCE BASED MEDICINE**

# TABLE OF CONTENTS

Table of Contents .....	2
Forward .....	6
Reference Materials .....	7
Textbooks .....	7
On-line Sources .....	7
Unit 1: Epidemiology Study Notes .....	8
Defining Epidemiology .....	8
The Importance of Epidemiology .....	8
Clinical (Symptoms & Signs) and Paraclinical Data .....	9
Harm, Diagnosis, Therapy, Prognosis, and Prevention .....	9
Disease Incidence .....	10
Cumulative Incidence .....	10
Incidence Density .....	11
Health of Canadians – Bronchogenic Carcinoma .....	11
Disease Prevalence .....	12
Point Prevalence .....	12
Period Prevalence .....	13
Health of Canadians – Diabetes Mellitus .....	13
Special Measures of Disease Frequency .....	13
Mortality Rate .....	13
Infant Mortality .....	13
Neonatal Death .....	14
Post-neonatal Death .....	14
Health of Canadians – Infant Mortality .....	14
Harm (Risk, Causality) .....	15
Measuring Risk (Harm) .....	16
Attributable Risk (AR) .....	17
Relative Risk (RR) .....	17
Population Attributable Risk (PAR) .....	17
Population Attributable Fraction (PAF) .....	18
Comparing Measures of Risk .....	18
Diagnostic Testing and Paraclinical Data .....	18
Gold Standard .....	19
Test Characteristics .....	20
Sensitivity and Specificity .....	22
Positive and Negative Predictive Values .....	24
Accuracy .....	25
Likelihood Ratios, Pre-test & Post-test Probabilities of Disease .....	25
Measures of Treatment Effect .....	27
Absolute Risk Reduction (ARR) .....	28
Relative Risk Reduction (RRR) .....	29

Relative Risk (RR).....	29
Odds Ratio .....	29
Number Needed to Treat (NNT).....	30
Efficacy and Effectiveness .....	32
Primary, Secondary, and Tertiary Prevention .....	32
Primary Prevention .....	32
Secondary Prevention .....	33
Tertiary Prevention .....	33
Measuring the Effects of Prevention .....	33
Natural History of Disease .....	33
Prognosis and Common Disease Outcomes .....	34
Health Related Quality of Life (HRQL).....	34
Health of Canadians – Life Expectancy.....	34
Prognostic Factors .....	35
Unit 2: Biostatistics Study Notes .....	36
Data .....	36
Nominal Data.....	36
Ordinal Data .....	36
Ranked Data .....	36
Discrete Data.....	37
Continuous Data.....	37
Measures of Central Tendency.....	37
Mean (average).....	37
Median.....	37
Mode .....	37
Example Calculations: Mean, Median, Mode .....	37
Using Measures of Central Tendency .....	38
Measures of Dispersion .....	39
Range.....	39
Variance .....	40
The Normal Distribution .....	40
Sampling Distribution and Standard Error of the Mean.....	42
Confidence Interval and Point Estimate .....	43
Independent and Dependent Variables.....	44
Data Presentation – Graphs .....	44
Box Plot.....	45
Survival Curve .....	45
Hypothesis Testing .....	46
Type I and Type II Error .....	48
Comparing Two Means from Independent Samples.....	48
Comparing Two Means from Paired Data.....	49
Comparing Proportions .....	49
Parametric and Nonparametric Tests of Statistical Hypothesis .....	49
Correlation .....	50
Regression Analysis .....	51
Sample Size Calculations .....	52

Unit 3: Research Design Study Notes .....	54
Quantitative Research .....	54
Observational versus Interventional Research .....	54
Correlational Study .....	55
Case Reports and Case Series .....	56
Cross-Sectional Surveys.....	56
Case-control Study .....	57
Cohort Study (Retrospective and Prospective).....	58
Intervention Studies (Clinical Trials) .....	59
Subject Selection.....	59
Subject Allocation.....	60
Levels of Blindedness (Concealment, Masking).....	60
Quasi-experimental Designs .....	61
Common Threats to the Research Validity and Reliability .....	61
Inference .....	61
Validity and Reliability .....	62
Bias .....	63
Intention to Treat Analysis versus Explanatory Analysis .....	64
Publication Bias.....	64
Confounding .....	64
Role of Chance.....	65
Qualitative Research.....	65
What is Qualitative Research? .....	65
Contrasting Quantitative and Qualitative Research.....	66
Assumptions Underlying Qualitative Research.....	66
Qualitative Research Methods .....	67
Qualitative Techniques for Data Collection.....	68
Sampling Strategies in Qualitative Research .....	68
Analyzing Qualitative Data .....	69
Maintaining Rigor in Qualitative Research.....	69
Conclusion.....	69
Unit 5: Evidence Based Medicine Study Notes .....	70
What is Evidence Based Medicine?.....	70
Formulating a Researchable Clinical Question .....	71
Searching for Evidence.....	71
Introduction to Critical Appraisal .....	72
Critical Appraisal – Articles about Therapy or Prevention .....	72
Critical Appraisal – Articles about Diagnostic Tests.....	74
Critical Appraisal – Articles about Prognosis .....	75
Critical Appraisal – Articles about Harm.....	76
Unit 6: Research Methods Study Notes .....	78
What is Clinical Research? .....	78
The Role of the Physician in Clinical Research .....	78
The Scientific Method .....	79
Introduction to the Research Proposal.....	79

Choosing a Topic for the Research Proposal .....	80
Title .....	81
Purpose .....	81
Background .....	81
Specific Aims .....	82
Study Design .....	82
Measurement .....	83
Data Handling and Analysis .....	83
Ethics .....	83
Budget .....	84
Significance and Relevance .....	84
Strengths and Weaknesses.....	84
Appendices.....	85

## **FORWARD**

With pleasure I distribute *Dr. Schaefer's Evidence Based Medicine*. Evidence Based Medicine is a paradigm that was created by Dr. David Sacket and colleagues. I'm grateful to Dr. Rob Haywood at the University of Alberta for having lead the development the Centre for Health Evidence website (cche.net) which I recommend to you. We are justifiably proud of the contributions of these Canadian physicians.

The objective of this document is to give students a concise overview and ready reference to assist with their studies and prepare for independent medical practice.

It is free for distribution and is usually available on my website:

<http://dr.schaferville.com>

I take credit only for gathering and distilling the ideas of others and hopefully I've presented it a way that is understandable. Credit for most of the graphics goes to Health Canada or documents that I've located on the internet.

I'm happy to receive comments or suggestions at: [jpschaef@ucalgary.ca](mailto:jpschaef@ucalgary.ca)

Regards,

Jeffrey P Schaefer MSc MD FRCPC

## REFERENCE MATERIALS

### Textbooks

Fletcher RH, Fletcher SW, Wagner EH. Clinical Epidemiology: The essentials, 3<sup>rd</sup> edition. Williams and Wilkins, Baltimore, 1996.

Hennekens CH, Buring JE. Epidemiology in Medicine. Little Brown and Company, Toronto, 1987.

Pagano M, Gauvreau K. Principles of Biostatistics. Duxbury Press, Belmont, 1993.

Sackett DL, Haynes BR, Guyatt GH, Tugwell P. Clinical Epidemiology. A basic science for clinical medicine. 2<sup>nd</sup> edition. Little, Brown and Company, Boston, 1991.

Sackett DL, Strains SE, Richardson WS, Rosenberg W, Haynes BR. Evidence-Based Medicine. How to Practice and Teach EBM. 2<sup>nd</sup> edition. Churchill Livingstone, Toronto, 2000.

### On-line Sources

Canadian Centre for Health Evidence

<http://www.cche.net/che/home.asp>

Canadian Institutes of Health Research: Tri-council Policy Statement for Ethical Research Involving Humans

<http://www.pre.ethics.gc.ca/english/policystatement/policystatement.cfm>

Sample Size Calculations at UCLS

<http://calculators.stat.ucla.edu/powercalc/>

Curriculum Information System

<http://cis01.medcis.net/portal/dt>

## **UNIT 1: EPIDEMIOLOGY STUDY NOTES**

The purpose of the epidemiology unit is to introduce the student to terms and concepts that are common to the study of health and disease. The student will also gain basic knowledge of the health status of Canadians.

### **Defining Epidemiology**

Epidemiology is the study of the distribution and determinants of disease frequency. There are two key assumptions in epidemiology: 1) human disease does not occur at random, and 2) human disease has causal and preventative factors that can be identified by systematic observation of populations or population subgroups.

The application of epidemiology to the bedside is often referred to as Clinical Epidemiology. Clinical Epidemiology has gained considerable importance since 1990 and related courses are relatively new phenomena in medical education.

### **The Importance of Epidemiology**

Physicians usually interact with only one patient at a time and the importance of observations derived from populations or population subgroups may not be readily apparent. However, physicians frequently use epidemiology, sometimes without being aware of it. For example, the clinical features of appendicitis have been described from centuries of observations of patients with this condition. Physicians encountering a patient with right lower quadrant abdominal pain will have considered the systematic observations of other physicians when considering appendicitis in such a patient. By applying the observations of others, physicians can diagnose common diseases and diseases they may have never encountered before.

In recent years, there has been much effort directed to refining and validating epidemiological information. With this effort, there has been a rediscovery of how powerful the tools of epidemiology can be when applied to the bedside. For example, physicians have been relying on the physical examination of the patient for centuries. Yet, only in the last decade has the usefulness of physical examination maneuvers been evaluated, sometimes with surprising results. Likewise, the usefulness of many time honored preventative and therapeutic interventions have been re-evaluated using the tools of epidemiology.

Recent developments in information technologies (computers, on-line access to data repositories) allow physicians immediate access to the latest medical information. Physicians in the office may access information sources during the patient encounter. However, accessible information does not always equate to good information. Physicians must know how to rapidly interpret the validity, results and applicability of medical knowledge. Critically appraisal of medical information is done using the tools of the epidemiologist.

## **Clinical (Symptoms & Signs) and Paraclinical Data**

It is important to be aware of the different types of data available to physicians. “Why are you seeing the doctor today?” is the time-honored question asked by those who greet our patients. The patient’s answer is known as the Chief Complaint.

Physicians need additional information before being able to resolve the chief complaint. Additional information comes from interviewing the patient and obtaining a Medical History. A medical history includes a description of the patient’s symptoms, past medical history, family history of health and disease, psychosocial history, history of medication use and intolerances or allergy, and a review of systems. Symptoms refer to the patient’s observations of their health status or simply, what they feel (physically or emotionally), see, hear, taste, or smell. “I see spots”, “I feel pain”, “I feel depressed” are examples of symptoms. The chief complaint is usually the symptom most important to the patient; hence, the special recognition.

Further information is obtained from examining the patient. From the physical or mental status examination, the physician obtains signs of disease and/or health. Signs are what the physician sees, smells, hears, feels, and tastes. Indeed, diabetes mellitus was diagnosed in ancient times by noting the patient’s urine tasted sweet! We probably don’t hug medical biochemists enough!

Symptoms and Signs are considered forms of Clinical Data.

In order to diagnose disease or monitor the effects of therapy, laboratory analyses of blood, urine, or other tissue is often required. Sometimes diagnostic imaging or other specialized investigations are needed as well. The results of such investigations are considered as Paraclinical Data.

## **Harm, Diagnosis, Therapy, Prognosis, and Prevention**

Most patient problems can be classified into five areas of inquiry.

Diagnosis: What disease is responsible for the abnormal findings?

Therapy: What therapy(s) (if any) is appropriate for a given disease?

Prognosis: What are the expected outcomes of a disease?

Prevention: How can disease be avoided or delayed?

Harm: What intervention or environmental agent may be contributing to disease?

Physicians must be able to identify the area to which a clinical problem belongs. Different problems are suited to different research methods. Different research methods are subject to different criteria for evaluating quality. Moreover, physicians are better able to anticipate the questions that are likely to be asked if they bear in mind these areas of inquiry.

Clinical Scenarios are useful for learning medicine because they simulate the clinical encounter. Some clinical scenarios lend to a single question while others may stimulate several questions for discussion. Consider the following clinical scenarios.

Clinical Scenario 1: A 55-year-old male presents to an emergency department complaining of chest pain that developed at lunch. It has been present for the last hour, is retrosternal, and radiates into the jaw. He has a history of hypertension and there is a family history of stroke. He smokes cigarettes, takes no medications, and has no allergy. Exam shows him to be distressed with normal vital signs. Cardiac exam reveals an S4.

What is the most important clinical question here? The answer is diagnosis! A physician cannot treat, prevent, or prognosticate if the disease has not been determined. This scenario presents clinical data that is most consistent with an unstable coronary syndrome (myocardial infarction or unstable angina). The differential diagnosis could include a foreign body (bone) in the esophagus.

Clinical Scenario 2: A 70-year-old female is found to have osteoporosis. She wants to know what should be done.

What is the most important clinical question here? In fact, more than one clinical question can be formulated from this scenario. We may wish to know how to treat her osteoporosis. More relevant to the patient, we may wish to know how to prevent outcomes commonly associated with osteoporosis such as hip and vertebral fractures.

Clinical Scenario 3: A 22-year-old male is treated for Hodgkin's lymphoma. He wants to know when he can return to work.

What is the clinical question here? He's interested in his prognosis. Prognosis may refer to any of several outcomes that may be associated with disease including disability.

Clinical Scenario 4: A 23-year-old women at 8 weeks gestation is concerned about the potential effect that working in front of a computer monitor may have on her unborn child.

What is the clinical question here? She is interested in harm. In the study of Evidence Based Medicine, harm usually refers to adverse outcomes associated with an exposure that is difficult to control.

### **Disease Incidence**

The incidence of disease (disease incidence) refers to new cases of disease among a population at risk for that disease over a specified time interval. This can be assessed by two methods.

#### Cumulative Incidence

Cumulative Incidence is the number of new cases of disease divided by the population at risk for that disease divided by the period over which all subjects were observed.

Cumulative Incidence = number of new cases / population at risk / time interval

It is common to index the cumulative incidence to a population of 10,000 (or 100,000) and period of 1 year. For example, according to data collected in 1998, the incidence of bronchogenic carcinoma (cancer of lung, bronchus, and trachea) in Canada was 6 per 10,000 per year. Therefore, another expression of cumulative incidence is:

Cumulative Incidence = number of new cases / 10,000 at risk / 1 year

### Incidence Density

Incidence Density is the number of new cases of disease divided by the total person-time of observation. Once again, the total person-time refers to persons 'at risk'.

Incidence Density = new cases / total person-time of observation

In studies of disease incidence, it may not be possible to follow all subjects for the duration of the study. Reasons include: migration, name change, cost, or withdrawal of consent to participate in the study. An Incidence Density can be calculated in this case.

Take the example of a study where 20,000 people at risk were to be followed for two years. However, it was only possible to follow 10,000 people for 2 years, 5,000 people for 1 year, and 5,000 people for 6 months. In this case there were (10,000 people x 2 yr) + (5,000 people x 1 yr) + (5,000 people x ½ year) = 20,000 + 5,000 + 2,500 = 27,500 person-years of observation. During the two-year study, there were 55 new cases of disease. The Incidence Density can be calculated as follows.

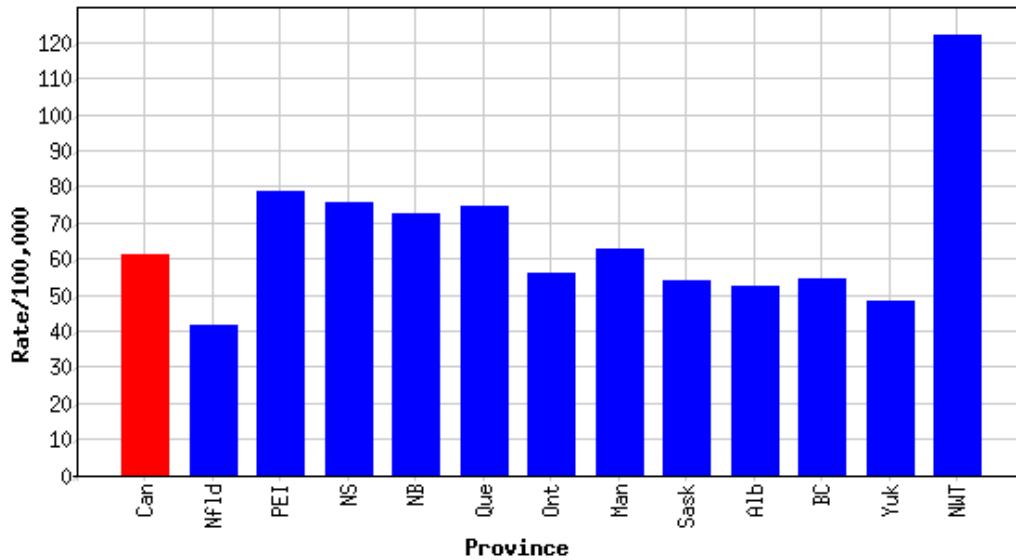
Incidence Density = 55 cases / 27,500 person-years  
= 0.002 cases / 1 person-year  
= 20 cases / 10,000 / year

A problem for those measuring the incidence of diseases by either method is verifying that the population used in the denominator is truly at risk. For example, in an incidence study of hepatitis A infection among Calgary high school students, it would not be valid to simply count the number of new infections because a proportion of students will be immune to hepatitis A owing to previous exposure or vaccination. In this case, testing for anti-hepatitis A antibodies would be required to evaluate risk status. Then the number of new cases among those at risk would be used to calculate an incidence. For diseases such as coronary artery disease or slowly growing cancers, establishing risk status can be challenging.

### Health of Canadians – Bronchogenic Carcinoma

Below is an example of the cumulative incidences of Cancer of the Lung, Trachea, and Bronchus, Both Sexes Combined, All Ages, 1998 from Statistics Canada by Province or

Territory. Time does not permit a discussion of age-standardization beyond indicating that this method accounts for differences in the distributions of age among the populations under consideration. Both a bar graph and table is provided.



Can	Nfld	PEI	NS	NB	Que	Ont	Man	Sask	Alb	BC	Yuk	NWT
61.10	41.74	78.57	75.54	72.69	74.72	55.83	62.72	54.05	52.30	54.62	48.30	122.25

Age-Standardized Incidence Rate per 100,000

### Disease Prevalence

Disease Prevalence refers to the proportion of individuals who have the disease among the observed population during a specified time interval (period). Although commonly said to be a rate, prevalence is really a proportion if you are a stickler for mathematical detail.

Prevalence = existing cases during a specified period / population under observation

#### Point Prevalence

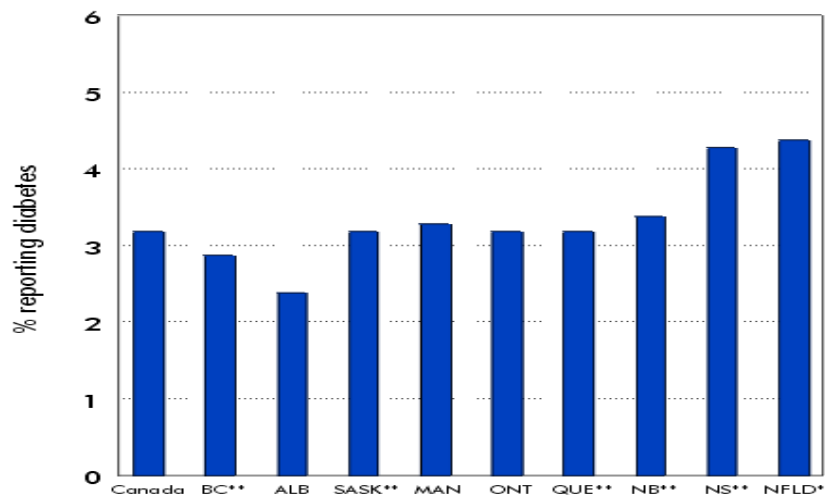
Point Prevalence refers to prevalence measures where the time interval is a 'point in time' or otherwise is very short. The point in time may refer to a point on the calendar. For example, 10% of all persons in the Day Care had diarrhea owing to Norwalk Virus on January 10, 2000. Alternatively, the point in time may be relative to a course of events that may vary from person to person. For example, 10% of all patients undergoing hip replacement surgery had low serum potassium on the third day after surgery.

## Period Prevalence

Period Prevalence refers to prevalence measures whereby the time interval is relatively long such as a month or year. This may be more appropriate for chronic diseases where disease status in the population is unlikely to change quickly. For example, 3.2% of Canadians had diabetes mellitus during the year ending April 30, 1997.

## Health of Canadians – Diabetes Mellitus

Below is the period prevalence rates of Diabetes Mellitus based on Self-Report for the Provinces during a Health Survey done during 1996 and 1997.



It is important to note the population under observation when interpreting disease prevalence. For example, a health economist may be interested in the prevalence of ovarian cancer among a population of males and females in order to evaluate the financial impact of this disease on the entire population. Alternatively, the Alberta Cancer Board may be interested in the prevalence of ovarian cancer among females alone in order to project the need for medical oncologists.

## **Special Measures of Disease Frequency**

Special measures of disease frequency are commonly used in medicine. Special measures are usually specific forms of disease incidence and prevalence.

### Mortality Rate

Mortality Rate is the same as incidence of death and may be limited to a specific disease or may be mortality from all causes.

### Infant Mortality

Infant Mortality refers to the death of a live-born infant within the first year of its life. Stillbirths (also referred to as fetal deaths) are not included in infant mortality. Infant mortality rates are usually based on the number of infant deaths per 1000 live births in

any given year. Sometimes, they are based on the number of infant deaths per 1000 population less than one year old. It should be noted that infant mortality “rates” are actually ratios, because infants who die in the year of interest, but were born in the previous year, are counted in the numerator but not in the denominator.

Neonatal Death

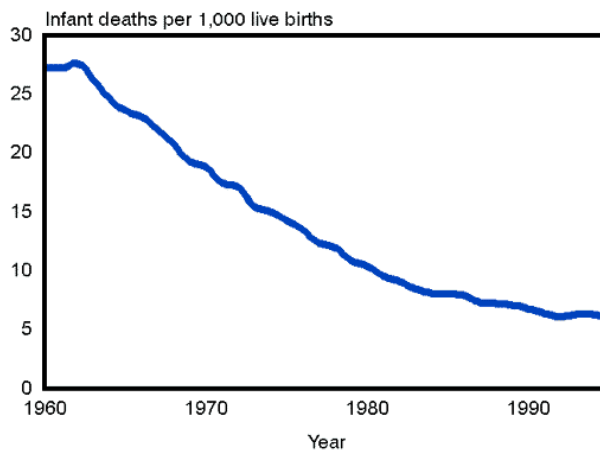
Neonatal Death refers to the death of an infant less than 28 days of age.

Post-neonatal Death

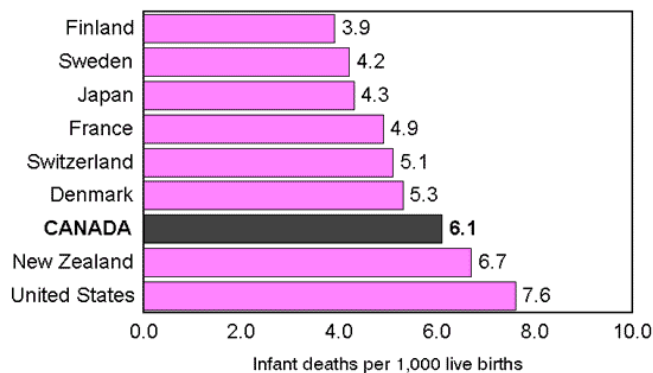
Post-neonatal Death refers to the death of an infant between 28 days and 1 year of age.

Health of Canadians – Infant Mortality

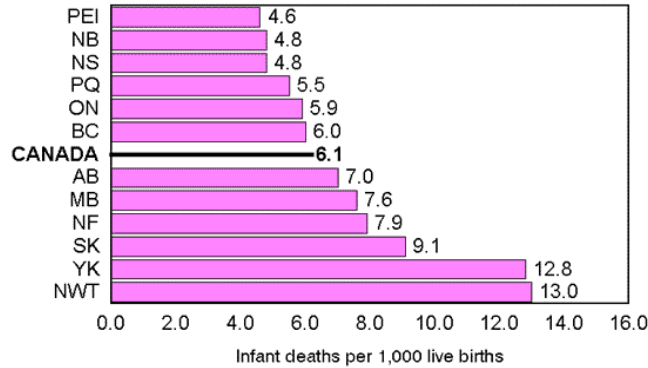
Below is data from Statistics Canada regarding special rates. Infant mortality rates are often used as an indicator of a country’s state of health development.



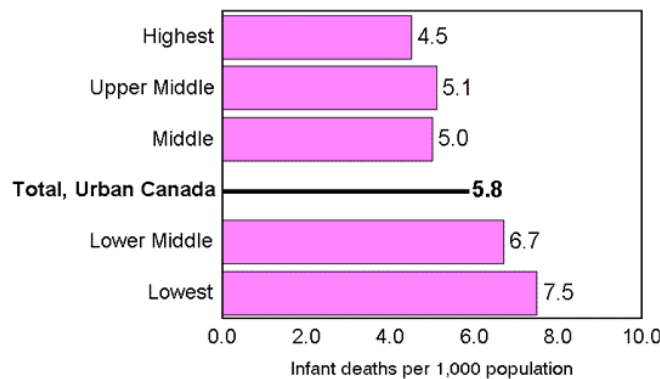
Infant Mortality Rate Canada, 1960-1995



Infant Mortality Rates Selected Countries, 1995



Infant Mortality Rates Provinces and Territories, Canada, 1995



Infant Mortality Rates by Income Quintile, Urban Canada, 1991

### Harm (Risk, Causality)

A cause is defined as ‘something that brings about an effect or a result’. Some diseases have a single cause whereas other diseases come about by a complex interplay of causes or risk factors. Among other things, epidemiologists concern themselves with propensity of a given exposure to cause disease. Such exposures are known as Risk Factors for disease. The relationship between a risk factor and disease is an issue of Harm.

Knowledge of what causes a disease may be crucial to understanding how to prevent, treat, diagnose, or provide a prognosis. Two sets of criteria for determining causation are worthy of review.

In 1882, Koch proposed criteria for establishing that a given disease was caused by a specific infectious agent using an animal model. Most texts refer to rabbits that he inoculated with *Streptococcus pneumoniae* bacteria (pneumococcus). Infected rabbits would get infected lungs (pneumonia) from which pneumococcus could be recovered and inoculated into healthy rabbits causing pneumonia once again from which pneumococcus could be recovered.

Koch’s Postulates (as they are called) included:

1. that the organism be present in every case of disease,
2. the organism be isolated from the diseased animal and grown in pure culture,
3. the organism causes the same disease when inoculated into a healthy animal, and
4. the organism must be recovered from that animal (step 3) and identified.

Until the widespread use of antimicrobials, infectious diseases were the very common causes of morbidity (illness) and mortality (death). Likewise, Koch's criteria for disease causality were important.

However, Koch's Postulates have three important limitations in the modern era. First, many diseases have multiple causes. For example, coronary artery disease is caused by smoking, diabetes, hypertension, and hyperlipidemia. For some individuals, no cause can be identified. Second, some exposures cause more than one disease. For example, smoking causes lung cancer, bladder cancer, emphysema, coronary artery disease, and peripheral arterial disease. Third, not all diseases are infectious.

In 1965, British statistician Sir Austin Bradford Hill proposed criteria for determining whether a causal relationship exists between an exposure and disease. These criteria are not specific to infectious diseases or any pathogenic mechanism. Not all criteria need be present in order to demonstrate causality. Rather, the more criteria that are present, the more likely that a given exposure is a risk factor for disease.

The criteria are (in random order):

Temporality: exposure precedes onset of disease

Strength: exposure strongly associated with disease frequency

Dose – Response: more exposure associates with higher disease frequency or severity

Reversibility: reduction in exposure associates with lower rates of disease

Consistency: association between exposure and disease is observed by different persons in different places during different circumstances

Biologic plausibility: causation is consistent with biological knowledge of the time

Specificity: one cause leads to one effect

Analogy: cause and effect relationship has been established for a similar risk factor or disease.

### **Measuring Risk (Harm)**

In this section, we consider different ways that the relationship between an exposure and disease may be expressed. Consider data from a study published in 1964 where a

population was followed for several years. At the beginning of the study, all members of the population were classified by smoking status (smoker, non-smoker). At the end of the study, deaths owing to lung cancer were counted among smokers and non-smokers.

Total death from lung cancer	0.56 / 1000 / year
Death rate from lung cancer among smokers	0.96 / 1000 / year
Death rate from lung cancer among non-smokers	0.07 / 1000 / year

Proportion of the population that smoked 0.56

A number of questions can be generated from this data.

### Attributable Risk (AR)

The Attributable Risk is the risk of disease experienced by those exposed to a given risk factor over and above that risk of disease experienced by those not exposed to the same risk factor. It is helpful to think of the average risk of a disease in a given population as the incidence of disease in that population. Therefore, the Attributable Risk is the difference in Incidence of Disease between the Exposed and Un-exposed populations.

Attributable Risk = (incidence among exposed) minus (incidence among non-exposed)

Attributable Risk = (0.96 / 1000 / year) - (0.07 / 1000 / year) = 0.89 / 1000 / year

Using our data, the incidence of fatal lung cancer among smokers that is over and above the incidence of fatal lung cancer among non-smokers is 0.89 deaths per 1000 per year. Alternatively, for every 10,000 smokers and 10,000 non-smokers, there are about 9 more fatal cases of lung cancer each year among the smokers.

### Relative Risk (RR)

The Relative Risk refers to how many times are exposed persons more likely to get the disease compared to non-exposed persons? The Relative Risk is sometimes referred to as the Risk Ratio. The relative risk is calculated by dividing disease incidence among those exposed by the disease incidence among those not exposed.

Relative Risk = (incidence among exposed) / (incidence among non-exposed)

Relative Risk = (0.96 / 1000 / year) / (0.07 / 1000 / year) = 13.7

Using our data, smokers are 13.7 times more likely to die from lung cancer compared to non-smokers.

### Population Attributable Risk (PAR)

The Population Attributable Risk refers to the contribution of a risk factor to the overall incidence of disease in the population (both exposed and non-exposed) under observation. The Population Attributable Risk is calculated by multiplying the

Attributable Risk by the Proportion of the Population Exposed.

Population Attributable Risk = (attributable risk) x (proportion of population exposed)

Population Attributable Risk = (0.89 / 1000 / year) x 0.56 = 0.50 / 1000 / year

Annually smoking accounts for 0.50 lung cancer deaths per every 1000 people in the population. Put another way, annually there are 5 lung cancer deaths attributable to smoking per every 10,000 people in the population.

#### Population Attributable Fraction (PAF)

The Population Attributable Fraction measures the proportion of disease that is attributable to exposure to the risk factor in the population. The Population Attributable Fraction is calculated by dividing the Population Attributable Risk by the Disease Incidence in the entire (exposed and non-exposed) population.

Population Attributable Fraction = (population attributable risk) / (population incidence)

Population Attributable Fraction = (0.50 / 1000 / year) / (0.56 / 1000 / year) = 0.89

From our data, 89% of the lung cancer deaths were accounted for by cigarette smoking.

#### Comparing Measures of Risk

There are advantages (and disadvantages) inherent to each measurement of risk. For example, the attributable risk is more meaningful at the level of an individual. That 9 extra people died of lung cancer, even among a relatively large population, can be sobering.

Relative risk can emphasize the strength of a causal relationship, particularly when the absolute value for disease incidence is small. In our example, the risk of dying from lung cancer among smokers was less than 1 in 1000. This low risk might not be very persuasive to a smoker who could counter that 999 out of 1,000 smokers did NOT die of lung cancer in this study. However, noting that smokers had nearly a 14 times higher risk of dying from lung cancer compared to non-smokers may serve as a better argument for quitting.

An advantage of the population attributable risk is that it allows us to compare the risk associated with a weak but highly prevalent exposure to a strong but rare exposure. From this type of analysis, we have learned that physicians can save more lives by treating everyone who has mild hypertension than treating only the few with severe hypertension.

#### **Diagnostic Testing and Paraclinical Data**

Making the correct diagnosis is one of the most important functions of a physician. Without a valid diagnosis, all other functions (therapy, harm, prevention, prognosis) are in jeopardy. Making a diagnosis is seldom a black and white affair. Rather, physicians

must deal with variable degrees of diagnostic uncertainty as they proceed onward from the patient's chief complaint.

Physicians rely heavily on the clinical information obtained from the history and physical examination. However, it is common to require certain investigations including blood analyses, diagnostic imaging, or histopathological examinations (to name a few) in order to confirm a diagnosis or rule out competing diagnoses. These investigations are commonly referred to as 'tests' by patients and their physicians. Some authors use the term, 'objective tests', to describe medical investigations but this can be misleading for two reasons.

First, medical investigations always require some human intervention for performance and interpretation. For example, a Magnetic Resonance Imaging (MRI) study is interpreted by a specialist in diagnostic imaging, the radiologist. Even if radiologists agree upon what is seen in the scan, they may differ on what the findings mean. Secondly, medical investigations usually produce a spectrum of results even when performed on healthy people. This spectrum may overlap with the results of investigations performed on people with disease. Thus, applying an arbitrary value to demarcate 'positive' and 'negative' test results may lead to diagnostic error.

In order to avoid the implications associated with the label, 'objective testing', Evidence Based Medicine Leaders may use the term, Paraclinical Data, when referring to the results of medical investigations. Because of its widespread use, we shall use the term 'diagnostic tests'. In this section, we will review the characteristics of diagnostic tests.

### **Gold Standard**

The degree to which a physician must be certain of a diagnosis will vary with the clinical scenario. For example, it is usually unnecessary to identify the species of virus responsible for an upper respiratory tract infection because interventions are marginally beneficial and this is a self-limiting condition. Alternatively, it is usual to make every attempt to diagnose the cause of rectal bleeding where the cause may include a surgically remediable cancer.

The Gold Standard refers to the method, procedure, or measurement that is widely accepted as being the best available to prove a given diagnosis in the clinical setting. In general Gold Standard tests are more costly, involved, and risky than other investigations so physicians and patients prefer to use less rigorous methods of evaluating a diagnosis, at least initially. When evaluating test performance, the gold standard is the diagnostic test to which other investigations are compared.

The diagnosis of pulmonary embolism (blood clot to the lung) is a good example the available diagnostic tests vary widely in cost, risk, and performance. Patients presenting to the emergency department with symptoms compatible with a pulmonary embolism may receive an inexpensive and rapid blood test to measure levels of D-

dimer. A negative test virtually excludes a pulmonary embolism; a positive test does not confirm the diagnosis so a more invasive test is required. A ventilation–perfusion scan is usually the next diagnostic test performed. For this diagnostic test, the patient receives an aerosolized and intravenous radiotracer. Afterward, the patient is placed under a radiation-detecting camera. A radiologist must interpret this study. The results may range from normal (no pulmonary embolism) to high probability (high likelihood of pulmonary embolism) with up to half the results being indeterminate (cannot rule in or rule out pulmonary embolism). In the absence of clear answer, the next diagnostic test is a pulmonary angiogram. In this procedure, a catheter is inserted into the femoral vein and threaded under x-ray guidance through the heart and into the pulmonary arteries where a radio-opaque dye is injected under a fluoroscope. The pulmonary angiogram is costly, invasive, and entails the highest patient risk; it is also the gold standard for diagnosing a pulmonary embolism.

As one can see from this example, there is a wide range of available tests. In the case of a pulmonary embolism, the Gold Standard is a pulmonary angiogram.

**Test Characteristics**

Many issues may be considered when selecting diagnostic tests. For many reasons, the gold standard may not be appropriate. How then do we evaluate diagnostic tests that are not considered Gold Standards?

One method is to create a table that compares the presence or absence of disease to the result of a diagnostic test. Consider the table below where the presence or absence of a pulmonary embolism (PE) (according to pulmonary angiogram, the Gold Standard) is compared to the results of ventilation-perfusion (V/Q) scanning in a study of 200 patients. A positive test suggests that disease is present and vice versa. What can we say about this study?

		DISEASE (PE)		Totals
		Present	Absent	
TEST (V/Q scan)	Positive	80	20	100
	Negative	10	90	100
Totals		90	110	200

There were 80 cases of PE where the V/Q scan was positive. These are True Positive V/Q Scans.

There were 10 cases of PE where the V/Q scan was negative. These are False Negative V/Q scans.

There were 20 cases of no PE where the V/Q scan was positive. These are False Positive V/Q scans.

There were 90 cases of no PE where the V/Q scan was negative. These are True Negative V/Q scans.

Our table can be updated to include these designations and an associated letter for each cell of the table.

		DISEASE (PE)		
		Present	Absent	
TEST (V/Q scan)	Positive	TRUE POSITIVES a = 80	FALSE POSITIVES b = 20	a + b = 100
	Negative	FALSE NEGATIVES c = 10	TRUE NEGATIVES d = 90	c + d = 100
		a + c = 90	b + d = 110	a+b+c+d = 200

The nomenclature T = test, D = disease, + is for positive / present, and – is for negative / absent is commonly used in textbooks. The following table can be presented.

Possible Results	Interpretation	Abbreviation	Mathematical
test is positive & disease is present	True Positive	T+ D+	a = 80
test is positive & disease is absent	False Positive	T+ D-	b = 20
test is negative & disease is present	False Negative	T- D+	c = 10
test is negative & disease is absent	True Negative	T- D-	d = 90

With this background, we can calculate the characteristics of the V/Q scan by

considering the probability of each outcome (true positive, false positive, true negative, false negative) given a test or disease outcome. 'P' is taken to mean probability.

Test Characteristic	Interpretation	Abbreviation	Mathematical
sensitivity	the probability that the test will be positive given that disease is present	$P(T+   D+)$	$= a / (a+c)$ $= 80 / (80+10)$ $= 0.888$ $= 88.9\%$
specificity	the probability that the test will be negative given that disease is absent	$P(T-   D-)$	$= d / (b+d)$ $= 90 / (20+90)$ $= 0.818$ $= 81.8\%$
positive predictive value	the probability that disease will be present given a positive test result	$P(D+   T+)$	$= a / (a+b)$ $= 80 / (80+20)$ $= 0.80$ $= 80\%$
negative predictive value	the probability that disease will be absent given a negative test result	$P(D-   T-)$	$= d / (c+d)$ $= 90 / (10+90)$ $= 0.90$ $= 90\%$
accuracy	the probability that the test result will be true	$\frac{P(T+ D+)+P(T- D-)}{2}$	$= (a+d)/(a+b+c+d)$ $= (80+90)/200$ $= 0.85$ $= 85\%$
positive likelihood ratio	ratio of: the probability of a positive test when disease is present AND the probability of positive test when disease is absent	$\frac{P(T+ D+)}{P(T+ D-)}$	$= (a/(a+c))/(b/(b+d))$ $= (80/90)/(20/110)$ $= 4.89$
negative likelihood ratio	ratio of: the probability of a negative test when disease is present AND the probability of negative test when disease is absent	$\frac{P(T- D+)}{P(T- D-)}$	$= (c/(a+c))/(d/(b+d))$ $= (10/90)/(90/110)$ $= 0.136$

### Sensitivity and Specificity

The test characteristics describe how well the V/Q scan performed in this hypothetical study. The sensitivity of 88.9% indicates that V/Q scanning was positive in 89 of 100 patients who had a PE in this study. The converse is also true. A physician relying only on the V/Q scan will fail to diagnose PE in 11 out of 100 patients who actually had a PE. Given that the mortality rate associated with PE varies around 15 – 20%, a false negative rate of 11% is considered unacceptably high.

Specificity may be similarly interpreting using the absence of PE in the denominator.

As important as sensitivity and specificity are to our understanding of a test, it can be argued that these characteristics are clinically irrelevant. Recall that the sensitivity (specificity) of a test is the probability that the test will be positive (negative) given the presence (absence) of disease. In reality, if disease status is known, there is no need to order the test!

Another drawback to sensitivity and specificity is that they assume there is a clear distinction between a normal and abnormal value for a test result which is not often the case. For example, concentrations of LDL cholesterol drawn in a population will return a range of values normally distributed about a population mean (bell shaped curve). It is not clear where abnormality may reside in this case and the test will be 'positive' or 'negative' based on where a cut point is selected. Given that high LDL cholesterol is a risk factor for heart disease, clinically significant hyperlipidemia will be under diagnosed (false negative test) if the cut point between normal and abnormal is set too high (vice versa for too low a cut point). A problem specific to LDL cholesterol is that this cut point has changed with each successive guideline published. Hence, the characteristic of the test has been changing with time.

Despite these drawbacks, sensitivity and specificity are invaluable to our ability to select the sequence in which tests are performed.

Imagine that you wish to set up a screening program for breast cancer. In this scenario we want to use a test that is VERY SENSITIVE. The more sensitive the test, the higher the probability that the test will return positive and the frequency of true positives test results (both true and false) will increase relative to negative test results.

The purpose of screening programs is to detect disease at an early stage when interventions are most likely to be beneficial. Screening programs are also intended to test large numbers of patients. Therefore, an inexpensive, non-invasive, and highly sensitive test is the best choice for disease screening. A mammogram is an example of such a diagnostic test and it is recommended for breast cancer screening.

The disadvantage of highly sensitive screening tests is the relatively high probability of having false positive results. To address this problem, physicians perform a highly specific confirmatory test before diagnosing the patient with breast cancer and starting therapy that is toxic and / or invasive. A breast biopsy is such a test and this is the gold standard for confirming breast cancer.

Clearly, a breast biopsy would be an inappropriate test for a population of women but when used selectively this invasive procedure provides the most good given the associated risk.

Another way to think about highly sensitive tests that are used in screening programs is that they rarely have false negatives. Hence, a negative screening test provides strong

reassurance that disease is not present. It is good for Ruling Out disease and the term 'SN-OUT' may serve as a memory aid.

Highly specific tests are used to confirm the presence of disease when it is suggested by clinical or other paraclinical data. The breast biopsy is such an example. Specific tests seldom give false positive results and are useful when the required therapy is toxic, emotionally traumatic, or expensive. Specific tests are good for ruling in disease and the term 'SP-IN' may provoke one's memory in an exam situation.

Positive and Negative Predictive Values

Recall that that positive predictive value in the V/Q scan study was 80%. In other words, for every 100 positive V/Q scans performed, there will be 80 patients with PE and 20 patients in whom PE was absent.

The negative predictive value was 90% in this study.

Positive and negative predictive values fit better with the physician's need to predict for the presence or absence of disease based on the knowable test result. However, the widespread use of these test characteristics is hindered by one major drawback; positive and negative predictive values vary with disease prevalence!

Take another look at the V/Q scan versus PE contingency tables. You'll note that the prevalence of disease was  $(a+c) / (a+b+c+d) = 90 / 200 = 0.45 = 45\%$ . What would the table look like if the prevalence were higher, say 55%? The table below provides the updated values. What happened to the predictive values? What happened to sensitivity and specificity?

		DISEASE (PE)		
		Present	Absent	
TEST (V/Q scan)	Positive	TRUE POSITIVE a = 80 97.7	FALSE POSITIVE b = 20 16.4	a + b = 114.1
	Negative	FALSE NEGATIVE c = 40 12.2	TRUE NEGATIVE d = 90 73.6	c + d = 85.8
		a + c = 90 110	b + d = 110 90	a+b+c+d = 200

Let's check:

prevalence =  $110 / 200 = 0.55 = 55\%$

sensitivity =  $97.7 / 110 = 88.8\%$  (unchanged from before)  
specificity =  $73.6 / 90 = 81.7\%$  (unchanged from before)

positive predictive value = 86.5% (was 80%)  
negative predictive value = 85.8% (was 90%)

As promised, the predictive values have changed with prevalence!

Another way to consider how this works is to imagine that a labeling error occurred and the blood of 100 females was tested for Prostate Specific Antigen (PSA, a screening test for prostate cancer). Imagine that an elevated PSA is found on one of these samples. Given that the prevalence of prostate cancer is 0% among females, it is intuitive (and mathematical) that the positive predictive value of this positive result will be 0%. This is considerably different from the positive predictive value of a PSA done on a male patient with a prostate nodule. In this case, the positive predictive value of a positive PSA for prostate cancer will be very high.

### Accuracy

Accuracy refers to the probability of obtaining a true result among all test results. This test characteristic is of limited utility because a highly specific test and a highly sensitive test may share the same accuracy. Therefore, the behavior of a test will be lost if accuracy alone is used describe its characteristics.

### Likelihood Ratios, Pre-test & Post-test Probabilities of Disease

In recent years, there has been increasing appreciation of the likelihood ratio in clinical medicine. Advantages of this test characteristic lay in its interpretation and application. A positive likelihood ratio expresses how many times more likely a positive test result will be found among those with disease when compared to those without disease. A negative likelihood ratio expresses how many times more likely a negative test result will be found among those with disease when compared to those without disease. (Negative likelihood ratios are usually less than 1.0.)

Likelihood ratios can be used to modify the physician's assessment of the probability (likelihood) of disease. You may not be familiar with likelihood ratios. How much do Likelihood Ratios (LRs) change disease likelihood?

LRs  $>10$  or  $<0.1$  cause large changes in likelihood.

LRs 5-10 or 0.1-0.2 cause moderate changes.

LRs 2-5 or 0.2-0.5 cause small changes.

LRs between  $<2$  and 0.5 cause little or no change.

As you can see, a diagnostic test that offers a likelihood ratio of 1.0 is useless. It offers no more diagnostic information that what was available to the physician prior to ordering the test. On the other hand, a test that offers a likelihood ratio greater than 10 will

strongly influence a physician toward making the associated diagnosis.

There are at least three advantages associated with the use of likelihood ratios in clinical medicine.

First, likelihood ratios can be developed for different values of a given test result. For example, an elevated serum lipase is associated with acute pancreatitis. The normal range for serum lipase is 0 – 60 units/litre. A patient may present with symptoms of acute pancreatitis and physicians may find the serum lipase to be 5,000 U/l or higher. In this situation, physicians can be quite confident that the result is abnormal and the patient has acute pancreatitis. However, another patient with similar symptoms may present with a serum lipase of 85 U/l. Knowing that healthy people may have laboratory results that are outside the normal range, a physician will be less confident that this result is truly abnormal and likewise, may not conclude that acute pancreatitis is present without additional supporting information. In fact, it is common for entirely asymptomatic and healthy people to have serum lipases as high as 85 U/l.

To help physicians deal with multi-level laboratory results, likelihood ratios can be determined for different levels of results. In this case, the positive likelihood ratio for acute pancreatitis associated with a serum lipase of 5,000 U/l may be 20. The positive likelihood ratio for acute pancreatitis associated with a serum lipase of 85 U/l may be 1.2. Instead of using intuition, physicians can use likelihood ratios to refine their diagnostic estimate.

Second, an advantage of likelihood ratios is that it combines sensitivity and specificity into one value that, unlike accuracy, conveys useful information. The interpretations of selected ranges of likelihood ratios are provided above.

Third, the likelihood ratio can also be applied to the pre-test probability of disease to produce a post-test probability of disease. Let's explore this further.

The probability of disease before ordering a given diagnostic test is known as the Pre-Test Probability of Disease. If patients are not selected, the pre-test probability of disease equals the prevalence of disease in the population. The Likelihood Ratio can be applied to the Pre-test Probability of Disease to produce a Post-test Probability of Disease.

Consider a female that undergoes a screening mammogram. Prior to undergoing the mammogram, the probability of her having breast cancer (in the absence of other information) is equal to the prevalence of breast cancer in the population. This is the pre-test probability of disease, which in her case is 10 per 1,000. Imagine that her mammogram shows her to have a 1 cm spiculated calcification and that this finding is associated with a likelihood ratio of 20 (a very suspicious lesion). With the information provided by the mammogram, we can calculate her probability of having breast cancer

according to the following equation.

Pretest odds x Likelihood Ratio = Posttest odds

There is some awkwardness associated with calculating odds. Odds is defined as the probability of an event divided by 1 minus the probability of an event or stated:

$$\text{Odds} = P(\text{event}) / (1 - P(\text{event}))$$

and

$$\text{Probability} = \text{Odds} / (1 + \text{Odds})$$

Going back to our patient:

prevalence of disease = pre-test probability of disease =  $10/1000 = 0.01$

pretest odds =  $0.01 / (1 - 0.01) = 0.01$

(note that odds and prevalence are similar for large or small numbers)

Post-test odds =  $0.01 \times 20 = 0.2$

Post-test probability =  $0.2 / (1 + 0.2) = 0.167$

Before the mammogram, our patient faced a 1% chance of disease. After the mammogram, our patient faces a 16.7% chance of disease. This is a considerable increase in the probability of disease. Given this result, she and her physician will want to consider a breast biopsy to confirm the impression of the radiologist who read the mammogram.

### **Measures of Treatment Effect**

The clinical trial is best source of evidence on which decisions about therapy and prevention are based. This is a relatively recent development in the history of medicine as clinical trials were largely unknown in the 1960s. The results of clinical trials can be expressed in different ways and it is important that physicians are able to interpret what is being presented.

Details of clinical trial design will be discussed elsewhere in the course. Briefly, the objective of a clinical trial is to compare the effects of one or more medical interventions. In the most common form of clinical trial, patients entering the trial are randomly allocated to either a control arm or treatment arm of the trial. Patients are observed for a time interval and outcomes of interest are measured at the end of a trial and compared.

Consider information derived from the Veterans Administration Trial in which 201 patients with hypertension were enrolled. Each subject was randomized to receive either a placebo (control arm) or an anti-hypertensive medication. After the period of observation was completed, the number of strokes, myocardial infarctions, and deaths

were summed in each group and compared. It was not surprising that the subjects provided with ant-hypertensive medication suffered fewer bad outcomes than the subjects who were provided with placebo. This reminds me of something Steven Wright once said, "I'm addicted to placebos. I'd give them up, but it wouldn't make a difference."

The investigators were also interested in the effects of anti-hypertensive medication on among subjects with and without target organ damage. Target organ damage refers to the adverse effects of hypertension on selected organs such as the heart, kidneys, and retina. Below are the study results showing the percentage of subjects that had any one of a stroke, myocardial infarction, or death stratified by the presence of target organ damage.

	<b>COMBINED STROKE, MI, DEATH (%)</b>	
	<b>PLACEBO</b>	<b>TREATMENT</b>
<b>DAMAGE</b>	22.2%	8.5%
<b>NO DAMAGE</b>	9.8%	4.0%

From 'eyeballing' the data, it appears that those with target organ damage seemed to have benefited more from therapy than those without target organ damage. By applying Measures of Treatment Effect, it is possible to express these differential effects explicitly.

As we proceed, it may be helpful to remember that risk of disease is the same as the probability of disease, which is the same as incidence of disease, which is the same as the event rate during the period of observation of the clinical trial.

Absolute Risk Reduction (ARR)

The Absolute Risk Reduction (ARR) is the difference between the probability of outcomes in the control and treatment arms of a clinical trial. "P(event)" refers to the probability of an event.

$$\text{Absolute Risk Reduction} = P(\text{outcome among controls}) - P(\text{outcome among treated})$$

$$\begin{aligned} \text{ARR with target organ damage} &= 0.222 - 0.085 &= 0.137 \\ \text{ARR without target organ damage:} &= 0.098 - 0.040 &= 0.058 \end{aligned}$$

From the Absolute Risk Reductions we see that those with target organ damage received a greater benefit than those without. However, a disadvantage associated with

the Absolute Risk Reduction is that the numbers are not intuitive. They do not easily convey a sense 'what's going on' and they are hard to remember.

### Relative Risk Reduction (RRR)

The Relative Risk Reduction is the absolute risk reduction divided by the probability of outcome in the control group (or at baseline).

Relative Risk Reduction = Absolute Risk Reduction / P(outcome among controls)

$$\text{RRR with target organ damage} = (0.222 - 0.085) / 0.222 = 0.62$$

$$\text{RRR without target organ damage:} = (0.098 - 0.040) / 0.098 = 0.59$$

From the Relative Risk Reduction we get the impression that there was little difference between the effects of hypertension therapy in our patient groups. The similarity of effect is almost surprising given our initial assessment of the table. A disadvantage of the Relative Risk Reduction is that it is sensitive to low probability outcomes in the control group. An analogy is that I can increase the chance of winning a lottery by 100% if I purchase a second lottery ticket.

### Relative Risk (RR)

The Relative Risk (RR) is the ratio of the probability of disease in the treatment group and the probability of disease in the control group.

Relative Risk = P(outcome among treated) / P (outcome among controls)

$$\text{RR with target organ damage:} = 0.085 / 0.222 = 0.38$$

$$\text{RR without target organ damage:} = 0.040 / 0.098 = 0.41$$

The Relative Risk imparts a comparison similar to that of the Relative Risk Reduction. It appears to minimize the differences in the effects of anti-hypertensive therapy on the two groups of subjects. Note that the relative risk reduction increases as the relative risk decreases.

### Odds Ratio

The Odds Ratio is a 'ratio of two odds'. Recall that the odds of an event equal the probability of an event divided by 1 minus the probability of the event.

$$\text{Odds of an Event} = P(\text{event}) / (1 - P(\text{event}))$$

Therefore, the odds ratio equals the odds of one event divided by the odds of another event.

$$\text{Odds Ratio} = \text{Odds (Event A)} / \text{Odds (Event B)}$$

Odds Ratios can be used to express the association between any two binary variables. Binary variables are 'yes / no' type variables and are discussed in the Biostatistics Unit. Odds Ratios are not intuitive unless one spends a lot of time at the racetrack. However,

they persist owing to some unique statistical properties. For example, logistic regression analysis uses odds ratios and it is relatively easy to calculate confidence intervals associated with odds ratios.

When Odds Ratios are used to measure a treatment effect, we need to determine the probability of the outcome associated with each treatment and then convert the probabilities into odds.

$$P(\text{outcome}) = (\text{number of outcomes among population}) / (\text{population})$$

$$\text{Odds (outcome)} = P(\text{outcome}) / (1 - P(\text{outcome}))$$

$$\text{Odds Ratio} = \text{Odds (outcome among treated)} / \text{Odds (outcome among controls)}$$

$$\begin{aligned} \text{OR with target organ damage:} &= [0.085 / (1 - 0.085)] / [0.222 / (1 - 0.222)] \\ &= [0.0929] / [0.285] \\ &= 0.33 \end{aligned}$$

$$\begin{aligned} \text{OR without target organ damage:} &= [0.040 / (1 - 0.040)] / [0.098 / (1 - 0.098)] \\ &= [0.0417] / [0.109] \\ &= 0.38 \end{aligned}$$

The Odds Ratio suggests a larger difference in between patient groups than the relative risk and relative risk reduction.

### Number Needed to Treat (NNT)

The Number Needed to Treat is the reciprocal of the absolute risk reduction. The advantage of this measure of treatment effect lay in its ability to convey information that is readily meaningful to physicians and patients. It also avoids the drawback of relative measures of treatment effect that are encountered when dealing with small numbers (see the lottery analogy – relative risk section).

The Number Needed to Treat refers to the number of subjects in a clinical trial that needed to be treated in order to prevent one outcome over the period of observation.

$$\text{Number Needed to Treat} = 1 / \text{Absolute Risk Reduction}$$

$$\begin{aligned} \text{NNT with target organ damage:} &= 1 / 0.137 &= 7 \\ \text{NNT with no target organ damage:} &= 1 / 0.058 &= 17 \end{aligned}$$

From the Number Needed to Treat we see that there is a clear difference in the treatment effect between the groups. Among those with target organ damage, 7 patients needed to be treated to prevent an outcome (any one of stroke, myocardial infarction, or death) over the duration of the trial. Among those without target organ damage, 17 patients needed to be treated to prevent an outcome over the duration of the trial.

The Number Needed to Treat is becoming the most common method of conveying trial results. To provide a sense of the values, the results of 6 clinical trials are presented below.

Therapy	Endpoint	NNT (5yr)
stepped care for diastolic BP of 115 – 129	death, stroke, myocardial infarction	3
coronary artery bypass grafting for left main disease	death	6
ASA for transient ischemic attack	death, stroke	6
cholestyramine for hypercholesterolemia	death, myocardial infarction	89
INH for inactive tuberculosis	active tuberculosis	96
stepped care for diastolic BP 90 – 109	death, stroke, myocardial infarction	141

Below is a summary table of measures of treatment effect.

Measure of Effect	Formula P(D+) = probability of disease	Outcome	
		Damage	No Damage
Absolute Risk Reduction	$ARR = P(D+_{control}) - P(D+_{treatment})$	0.137	0.058
Relative Risk Reduction	$RRR = \frac{P(D+_{control}) - P(D+_{treatment})}{P(D+_{control})}$	0.62	0.59
Relative Risk	$RR = \frac{P(D+_{treatment})}{P(D+_{control})}$	0.38	0.41
Odds Ratio	$OR = \frac{P(D+_{treatment}) / (1 - P(D+_{treatment}))}{P(D+_{control}) / (1 - P(D+_{control}))}$	0.33	0.38
Number Needed to Treat	$NNT = \frac{1}{ARR}$	7	17

## **Efficacy and Effectiveness**

You may notice that the term 'efficacy' rather than 'effectiveness' is used to describe the effects of an intervention under the condition of a clinical trial. Briefly, treatment efficacy refers to the effect of treatment under ideal conditions, which is the usual condition of a clinical trial. Patients entering clinical trials are carefully screened for inclusion and exclusion criteria and significant effort is made to monitor the application of the intervention and the patient's progress and safety. Moreover, in clinical trials subjects agree to adhere to their allocated treatment as a condition of entry into the trial (although they are allowed to withdraw if they wish).

In the day-to-day practice of medicine, resources for such efforts are not usually available and patients are not usually excluded from receiving therapy owing to narrow criteria set forth in clinical trials. Moreover, patients are entirely free to accept or decline recommended interventions. In contrast to the ideal conditions of a clinical trial, effectiveness refers to how well an intervention performs in the 'real world'. Broadly speaking, a treatment is effective if it does more good than harm in those whom it is offered.

## **Primary, Secondary, and Tertiary Prevention**

Prevention is defined as "the act of keeping from happening". In medicine, we may intervene to prevent disease from occurring, delay the onset of disease, arrest the progression of disease, and/or slow the progression of disease. Physicians and patients usually think of prevention as being separate from therapy. However, this is a matter of perspective. What may be prevention to one is treatment to another.

For example, treating hypercholesterolemia may be thought of as preventing coronary artery disease. Treating coronary artery disease may be thought of as preventing a myocardial infarction. Treating a myocardial infarction may be thought of as preventing heart failure. Treating congestive heart failure may be thought of as preventing disability and death.

In this example, many physicians could be involved including the family physician, general internist, endocrinologist, cardiologist, and cardiac surgeon. Each may have a different perspective of how prevention applies to their patient. Understanding different types of prevention will be impossible unless the risk factor and the disease are defined explicitly and precisely.

Three levels of prevention are recognized: primary, secondary, and tertiary prevention. Unfortunately, there is some variability regarding the interpretation of secondary and tertiary prevention in the medical literature.

### Primary Prevention

Primary prevention refers to interventions designed to reduce the risk of disease onset. Examples of primary prevention include: vaccinating against influenza, prescribing cholesterol lowering medication to prevent myocardial infarction (in those without coronary artery disease), and educating new parents about car seats.

### Secondary Prevention

Secondary prevention refers to interventions designed to interrupt or minimize the progression of pre-clinical disease. Pre-clinical disease in this context means disease that has not progressed enough to cause symptoms. Screening programs are part of secondary prevention activities. For many conditions, detection at an early stage allows for interventions that are more effective, less invasive or risky, and/or less costly. For example, screening colonoscopy may detect potentially malignant polyps of the colon. Most polyps can be resected with the colonoscope (the operator simply snares the polyp with a wire). This is immeasurably preferable to waiting until symptoms of colon cancer develop before intervening with major surgery and/or chemotherapy.

### Tertiary Prevention

Tertiary prevention refers to interventions designed to slow or arrest the progression of clinically apparent disease. Clinically apparent disease means diseases for which symptoms and/or signs have developed. Put another way, tertiary prevention reduces the complications of known disease. An example of tertiary prevention would be referring a patient known to have diabetes mellitus for an ophthalmologic assessment for the detection and treatment of diabetic retinal disease. In this case, the tertiary prevention is directed toward the complications of diabetes mellitus. Again, prevention can be a matter of perspective. From the perspective of the ophthalmologist, the detection of diabetic eye disease prior to the onset of visual symptoms could validly be considered secondary prevention.

### **Measuring the Effects of Prevention**

Because of the parallels between prevention and treatment, we may use the measures of treatment effect when expressing the effects of a preventative intervention. Absolute risk reduction, relative risk reduction, relative risk, odds ratios, and number needed to treat (prevent) can be used.

### **Natural History of Disease**

The Natural History of Disease refers to the course disease in the absence of intervention. Knowledge of the natural history of disease is vital to making prognoses and evaluating the effects of preventative and/or therapeutic interventions. For example, many diseases are self-limiting. That is to say, they will resolve without intervention. Although not commonly used, the course of a disease when interventions are applied may be referred to as the Clinical History of Disease.

## **Prognosis and Common Disease Outcomes**

Prognosis refers to making predictions about the outcomes of disease. Because diseases may have limitless effects on humans, virtually any outcome may be considered. A number of outcomes are used frequently enough to merit definition.

5-year survival is the percent of patients surviving for 5 years from some point in their disease. This is commonly used in the field of oncology.

Case Fatality Rate is the percent of patients with a disease who die of it.

Disease-specific Mortality is the number of people per 10,000 (or 100,000) population dying of a specific disease.

Response is the percent of patients showing some evidence of improvement following an intervention.

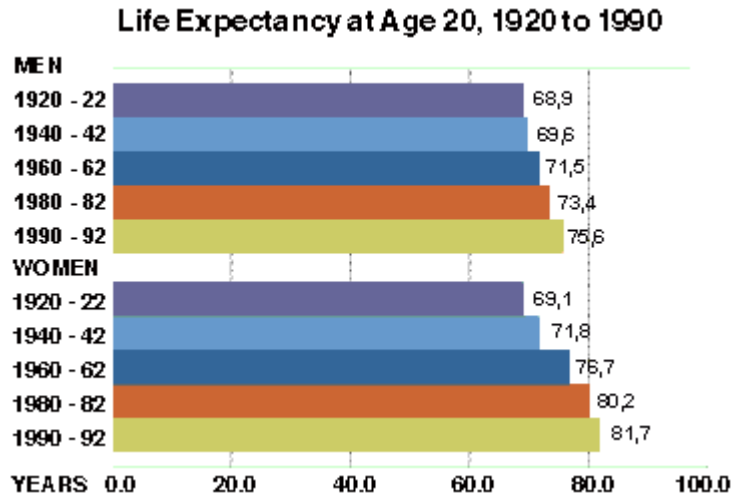
Remission is the percent of patients entering a phase in which disease is no longer detectable.

Recurrence is the percent of patients who have return of disease after a disease-free interval.

## **Health Related Quality of Life (HRQL)**

Thus far, we have primarily considered outcomes that refer to rates of mortality (death) or morbidity (non-fatal disease). However, how our patients feel and function are outcomes that have gained increasing attention over the last decade. Health Related Quality of Life (HRQL) refers to how our patients feel and function. HRQL is much broader in scope than mortality and morbidity. It may include virtually any dimension where disease has an impact on daily living. HRQL will probably receive increasing attention for at least two reasons. First, medical researchers have developed and continue to develop tools (surveys, questionnaires and other assessments) that can measure quality of life outcomes. Secondly, it may not be possible to substantially reduce rates of morbidity and mortality for many conditions in the absence of a new technology or scientific paradigm. Below is a table showing changes in Canadian life expectancy over the last century. Whether we will increase life expectancy by another 10 years is a matter of controversy. However, there may be tremendous opportunity to improve on quality of life in situations where mortality or morbidity has been optimized.

### Health of Canadians – Life Expectancy



Source - Statistics Canada

### Prognostic Factors

Something that is associated with a disease outcome is known as a Prognostic Factor. Prognostic factors can be positive or negative depending on the effect. For example, low urine output in a burn patient is associated with a higher frequency of kidney failure and subsequent death in this patient population. Therefore, it is a negative prognostic factor. Age is a commonly considered prognostic factor.

It is important to distinguish between a prognostic factor and risk factor. Risk factors for disease operate prior to the onset of disease while prognostic factors operate once a disease is present. Two interesting examples arise from cardiology.

A low blood pressure compared to high blood pressure is associated with a reduced risk of having a future myocardial infarction (MI, heart attack). However, having low blood pressure when compared to high blood pressure after the onset of an MI is associated with a higher mortality rate. The latter is presumably due to the presence of enough myocardial damage to impair the pumping action of the heart.

Another example relates to gender. At least prior to menopause, females are less likely than males to have an MI. However, females are more likely to die from an MI than males. Hence being female protects against the risk of MI but is negative prognostic factor.

## UNIT 2: BIostatISTICS STUDY NOTES

The purpose of the biostatistics unit is for the student to learn how to interpret the results of published medical research and for the student to have a basic understanding of the methods used to analyze data. It is acknowledged that there is variation among authors regarding some of the information presented. This section is based primarily on Principles of Biostatistics authored by Marcello Pagano and Kimberlee Gauvreau in 1993. For the purposes of examination, information in the core document shall be considered correct.

### Data

Every study or experiment yields a set of data. Qualitative Research tends to explore questions of 'what,' 'how,' and 'why.' The data associated with these questions tend to be narrative accounts, explanations, typologies, or conceptual frameworks. Quantitative Research tends to explore questions of 'whether' or 'how much.' Answers to these questions can be expressed numerically.

Numerical data may be characterized according to the following types.

#### Nominal Data

Nominal data is the simplest form of data. The purpose of nominal data is to differentiate one object from another. Some examples of Nominal Data include: categories of gender (male, female) or diseases (pneumonia, asthma, emphysema). Note that nominal data may be binary (dichotomous) which means the variable can be one of two values (gender) or there may be several values that the variable may take (diseases). Numbers can be assigned to categories in a nominal scale of measurement such as 1 = pneumonia, 2 = asthma, 3 = emphysema. However, there is no rationale for this particular ordering of these diseases (any other order would serve as well). Moreover, calculations using these numbers usually have no meaning. For example, an average disease of 2.5 is meaningless.

#### Ordinal Data

Ordinal data has the property of order. For example, injuries may be classified according to level of severity where '1' = fatal, '2' = severe, '3' = moderate, and '4' = minor. A natural order exists here. However, the differences between values are not important. The difference between 1 (fatal) and 2 (severe) may not be same as the difference between 2 (severe) and 3 (moderate) in our example. Therefore, it may not be appropriate to make calculations based on this form of data.

#### Ranked Data

Ranked Data is a special form of data that can be placed within an Ordinal Scale of measurement. Ranked data refers to the situation where observations are first

arranged according to some attribute after which a numerical label is applied. For example, we could list the causes of death in Alberta for the last year and then rank each cause according to frequency. The most common cause of death would be at position one, the second most common cause of death at position 2, and so on. It is possible to perform some statistical tests on ranked data.

### Discrete Data

Discrete Data has the property of both order and magnitude of change. The numbers represent measured quantities rather than serve merely as labels. Also, the difference between adjacent values is constant. Interval data is restricted to integers or counts such as the number of motor vehicle collisions in Calgary or number of admissions to hospital in a year. Arithmetic operations applied to this data usually result meaningful information although the results may not be an integer. For example, that average number of children per family is 2.4 is meaningful even if no family has this number of children.

Temperature and dates are also examples of interval data. This is a good example, where the difference between values is meaningful whereas ratios of the values are not. Twenty degrees Celsius is not twice as hot as 10 degrees Celsius!

### Continuous Data

Data that represent measurable quantities but are not restricted to integers or pre-specified values is known as Continuous Data. Examples include serum cholesterol values (e.g. 5.23, 6.33, 4.45 mmol/l) or infant weights (6,666, 5,423, 7,092 grams). Arithmetical operations can be applied to this data with meaningful outcomes.

## **Measures of Central Tendency**

It is useful to know the central tendency associated with a set of observations. There are three commonly used measures of central tendency.

### Mean (average)

The mean value is calculated by dividing the sum of the values in a data set by the number of observations.

### Median

The median is defined as the 50<sup>th</sup> percentile within a set of ranked data. If we arranged a list of observations from largest to smallest, then half the values would be greater than the median value and half the values would be less than the median value. If the number of values is even, we average the two middle values to obtain the median.

### Mode

The mode is the most frequently reported value in a data set.

### Example Calculations: Mean, Median, Mode

Consider a series of pulmonary function tests taken from 9 volunteers. The FEV 1 refers to the volume of air expired in the first second during a forced expiratory maneuver.

<b>Subject</b>	1	2	3	4	5	6	7	8	9
<b>FEV1</b>	2.2	3.5	2.6	2.8	2.8	2.3	2.2	2.8	2.8

number of observations = 9  
 sum of observations = 23.6  
 mean = 2.62 liters

The mean is calculated by dividing the sum (23.6) by the number of observations (9). Let's rank the data from lowest to highest so that we may select the median.

<b>Subject</b>	1	7	6	3	4	5	8	9	2
<b>FEV1</b>	2.2	2.2	2.3	2.4	2.6	2.8	2.8	2.8	3.5

median = 2.6

Which value is observed most frequently? The value '2.8' is observed 3 times; all other values are observed less frequently. The mode is 2.8.

Using Measures of Central Tendency

Why use one measure rather than another? One important reason lays in the sensitivity of the measure to extreme values. An example is length of stay data taken from hospital admissions. It is frequent for a few patients to need a long hospitalization. For instance, a patient's acute illness may have resolved but the need for lengthy rehabilitation may prevent return to the community.

Consider the length of stay (days) associated with Physician A and Physician B for 9 patients they admitted to hospital last fall.

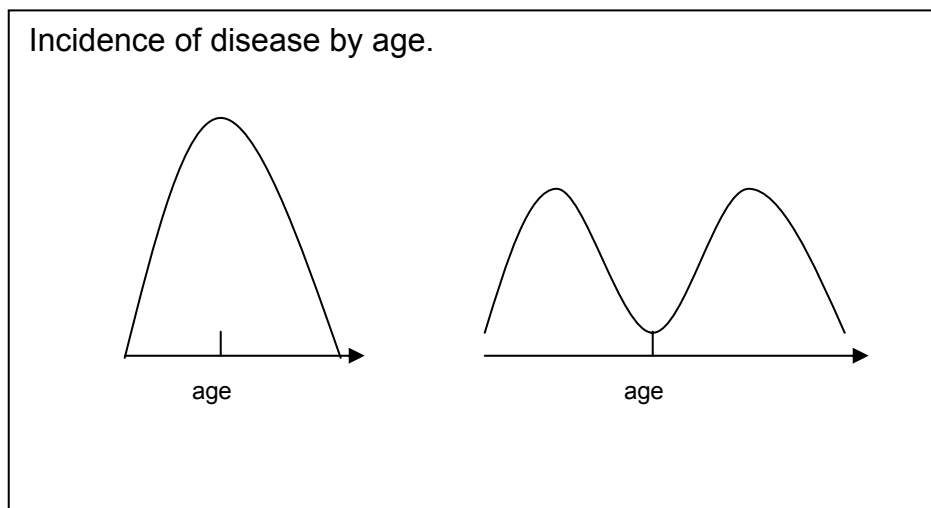
<b>LOS A (d)</b>	2	3	5	6	7	9	12	13	23
<b>LOS B (d)</b>	2	3	5	6	7	9	12	13	14

Mean LOS – physician A: 8.9  
 Mean LOS – physician B: 7.9  
 Median LOS – physician A: 7

Median LOS – physician B: 7

If we evaluated the performance of these physicians according to mean length of stay, we would question why patients cared for by Physician A required an additional day of hospitalization on average relative to Physician B. The reason for this, of course, was that one patient was unusually sick and the extreme length of stay influenced the mean value. The median did not change owing to the difference in LOS between one patient. The median is less sensitive to extreme values and is said to be more 'Robust' in that respect.

Another reason to consider measures of central tendency other than the mean is to better characterize data that may not be distributed around one value. Consider the distributions of disease incidences according to age as shown below. The mean and median values are the same for both distributions. The distribution on the left is unimodal and on the right is bimodal.



In the case of Hodgkin's Lymphoma, there are two peaks of incidence, age 29 years and age 73 years. Given this bimodal distribution, it was hypothesized that this disease might consist of two separate entities. Indeed, it was found that Hodgkin's Disease in young people behaves more like an inflammatory process where as in the elderly it behaves more like a neoplastic process. Had it not been for an assessment of mode, this feature of Hodgkin's Lymphoma may not have been appreciated.

### **Measures of Dispersion**

It is often of interest to know how the values within a dataset vary. Measures of Dispersion allow us to summarize the variability.

#### Range

The Range refers to the difference between the largest and smallest observations.

Let's reconsider the length of stay data for physicians A and B.

LOS A (d)	2	3	5	6	7	9	12	13	23
LOS B (d)	2	3	5	6	7	9	12	13	14

The range values for physicians A is:  $23 - 2 = 21$  days.

The range values for physicians B is:  $14 - 2 = 12$  days.

You will note that the range, like the mean, is sensitive to extreme values. Although the length of stays for 8 of 9 patients was the same, the ranges differed considerably.

### Variance

The Variance may be used to measure the dispersion of values in a set of data. The Variance is calculated according to the following steps using the length of stay data from our physicians.

Step 1: subtract means from each LOS to obtain difference

Step 2: square the differences

Step 3: sum the squared differences

Step 4: divide the sum of the squares by 8 (9 values - 1) to get variance

<b>LOS A</b>	2	3	5	6	7	9	12	13	23
mean	8.89	8.89	8.89	8.89	8.89	8.89	8.89	8.89	8.89
difference	-6.89	-5.89	-3.89	-2.89	-1.89	0.11	3.11	4.11	14.1
difference <sup>2</sup>	47.5	34.7	15.1	8.35	3.57	0.01	9.68	16.9	199

sum of diff <sup>2</sup>	334.89
n-1	8
<b>variance</b>	<b>41.86</b>
<b>std dev</b>	<b>6.47</b>

<b>LOS B</b>	2	3	5	6	7	9	12	13	14
mean	7.89	7.89	7.89	7.89	7.89	7.89	7.89	7.89	7.89
difference	-5.89	-4.89	-2.89	-1.89	-0.89	1.11	4.11	5.11	6.11
difference <sup>2</sup>	34.7	23.9	8.35	3.57	0.79	1.23	16.9	26.1	37.3

sum of diff <sup>2</sup>	152.89
n-1	8
<b>variance</b>	<b>19.11</b>
<b>std dev</b>	<b>4.37</b>

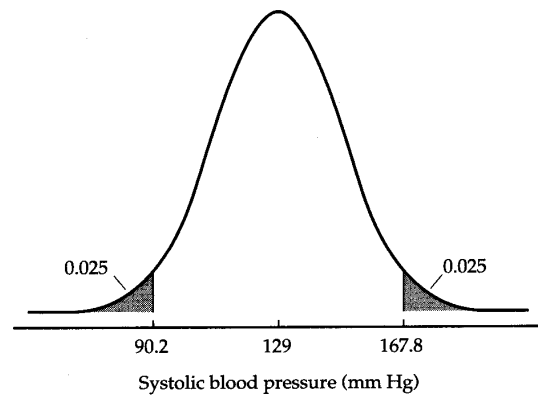
The Standard Deviation is simply the square root of the variance.

From our length of stay data, we see that the standard deviation is less different between the two physicians than the ranges had suggested.

### **The Normal Distribution**

Any characteristic that can be measured or categorized is known as a variable. In medicine, it is common for continuous data to distribute in a manner shown in the graph below that depicts blood pressure results. For the time being, ignore the references to

either end of the curve (the tails) but note that the mean, median, and mode blood pressure is 129 mm Hg.

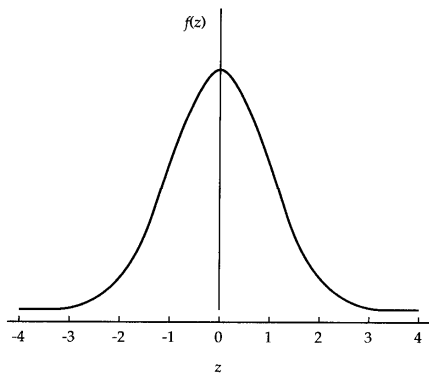


When the observations for a set of values appear like this graph, we may describe the data as Normal, or Gaussian, or bell-shaped.

Although there is a precise mathematical definition for the normal distribution, we generally are satisfied that data is normal if:

- 1) the curve shows that the highest frequency falls in the centre,
- 2) the curve (bell curve) is symmetrical so that the mean, mode, and median will coincide at the centre of the curve, and
- 3) the further away any particular value is from the average (above or below), the less frequent that value will be.

In a Standard Normal Curve, the mean is equal to 0, the mode is equal to 0, and the median is equal to 0 with a standard deviation of 1. A graphic example is shown below. Given these values, we can make some predictions about the dataset that fits a normal curve. For example, 68% of the area of the curve lays within the 1 standard deviation on either side of the mean and 95% of the area of the curve lays within 2 standard deviations on either side of the mean. Considering the blood pressure data above, you will note that 95% of the values for blood pressure are within two standard deviations of the mean blood pressure reading. Those with a blood pressure reading greater than 2 standard deviations from the mean would be considered to have an abnormal reading and this may be a sign of disease.

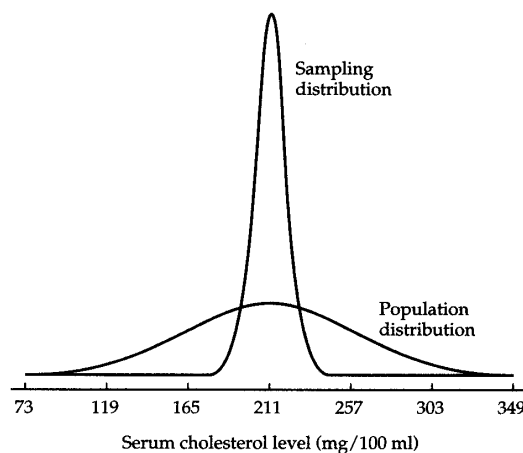


The standard normal curve for which  $\mu = 0$  and  $\sigma = 1$

### Sampling Distribution and Standard Error of the Mean

If we were interested in knowing the mean serum cholesterol value for a population we could either: 1) measure the serum cholesterol for every member of the population (and break our budget), or 2) estimate the mean by measuring the serum cholesterol in a sample taken from population by selecting members of the population at random.

Suppose that the population distribution for cholesterol was available to us for a small town (like Framingham, Mass) and that we took a random sample of 25 people and measured their serum cholesterol and calculated a mean value for that sample. After returning our sample back to the population at large, we repeated endlessly. Soon we would have a have enough sample means to plot against the background of the population distribution for cholesterol. The resulting graph would appear below.



Some very interesting features emerge. First, the mean, median, and mode of the sampling means would be the same population mean. Second, the dispersion of the

sampling means will be less than the population distribution. We can also calculate the standard deviation of the sample means to produce a number called the Standard Error of the Mean. The standard error of the mean is a very useful number because it forms an important component of the calculation of a Confidence Interval.

### **Confidence Interval and Point Estimate**

The confidence interval refers to a range of numerical values where we expect to find the true value with a specified degree of certainty. The point estimate refers to the calculated value associated with an outcome. For example, consider the outcome of a trial looking at a drug used to treat heart failure. In this trial, it was published that patients allocated to receive spironolactone compared to those receiving placebo had a relative risk for cardiac death or hospitalization of 0.68 (95%CI: 0.59 – 0.78).

Stepping back somewhat, these investigators were interested in knowing what effect spironolactone has on the incidence of cardiac death or hospitalization. One way to answer this question would be to gather the entire population of people who have heart failure and enter them into a randomized placebo controlled trial. Of course, this is impossible because there might be millions of people with heart failure on planet earth. The cost and logistics of such an endeavor would be prohibitive. However, the result of such an experiment would give us the true answer regarding the effect of spironolactone on this population.

The next best thing would be to sample the population of people with heart failure, which is what they did. In fact, 1,663 patients with heart failure were entered into the trial and randomized to receive either spironolactone or placebo. Cardiac death and hospitalization was observed. The relative risk for this outcome was 0.68 (95%CI: 0.59 – 0.78). What does this mean?

First, the point estimate of the relative risk was 0.68. The point estimate is the result we obtained based on the data in this trial knowing that if we repeated the trial again, we would have obtained a different outcome for relative risk owing to the play of chance. If we repeated the experiment an infinite number of times, we would have an infinite number of values for relative risk and if plotted these values they would be normally distributed just like the sample distribution shown in the previous section. Hence, the point estimate of 0.68 is one value among an infinite number of values that could have been obtained under the circumstances of this trial. So, how sure are we that our value of 0.68 is close to the population value? The answer to this question is the confidence interval (CI).

Recall that the published outcome for relative risk was 0.68 (95%CI: 0.59 – 0.78). The 95%CI of 0.59 – 0.78 means that true population relative risk will be within this range of values 95% of the time. Other ranges could have been reported such as the 90% confidence interval or the 99% confidence interval but it is usual to select the 95%

confidence interval.

The calculation of the confidence interval varies somewhat with the type of data being used, but one example of a calculation of a 95% confidence interval would be the point estimate plus and minus 1.96 multiplied by the standard error of the mean.

95% confidence interval = point estimate +/- (1.96 x standard error of the mean)

If we were interested in a 99% confidence interval, the value 1.96 is replaced by the value 2.58). The standard error of the mean is really a standard deviation of sample means and this number will decrease when the 'n' (sample size) increases. An important implication of this is that the confidence interval will narrow when more people are observed. That is to say, the larger the sample size, the narrower our confidence interval.

### **Independent and Dependent Variables**

Quantitative research often seeks to find a relationship between two or more variables. When embarking on an area of inquiry it is useful to define the independent and dependent variables to be examined. The independent variables are those that are felt to influence the outcome under consideration. Dependent variables usually refer to the outcome under consideration. For example, in a study assessing the relationship between diastolic blood pressure and annual stroke incidence, blood pressure would be considered the independent variable and stroke the dependent ('it depends on blood pressure') variable. Although this is a useful framework, it should be noted that a strong mathematical relationship (association) between variables is not proof of causality in itself. For example, smoking is known to cause lung cancer and smokers consume more alcohol than non-smokers. Given these facts, an association will be found between alcohol consumption and lung cancer even though there is no causal relationship between these two variables. This is an example of confounding.

Variables that are dependent / independent may also be labeled as output / input variables or response / explanatory variables. It is usual to graph the dependent variable on the y-axis and the independent variable on the x-axis in graphs.

Lastly, for any assessment of an association, it is possible (and common) to assess many independent variables but only one dependent variable. For example, in determining risk of liver failure among those with hepatitis C infection, researchers are likely to measure the relative contributions of age, gender, alcohol use, and other co-existing liver diseases.

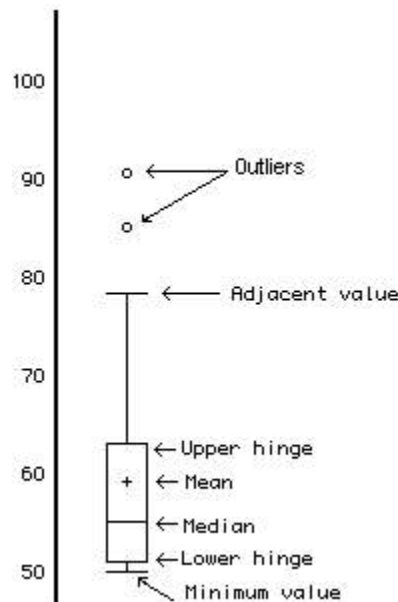
### **Data Presentation – Graphs**

Graphs allow the reader to assess general patterns of data within a dataset at a glance. Several types of graphs exist including bar graphs, frequency histograms, box-plots, and line-graphs. You are likely to be familiar with common forms of graphs. However,

two types of graphs merit description.

### Box Plot

A box plot provides an excellent visual summary of a distribution. The box stretches from the lower hinge (defined as the 25th percentile) to the upper hinge (the 75th percentile) and therefore contains the middle half of the values in the distribution (dataset). The median is shown as a line across the box. Therefore, 1/4 of the distribution is between this line and the top of the box and 1/4 of the distribution is between this line and the bottom of the box. Sometimes, the mean is marked with a point or other symbol. The lines (or 'whiskers') projecting out from the box extend to the adjacent values. Adjacent values are the most extreme values that are not more than 1½ times the height of the box beyond either quartile. All points outside this range are considered outliers.

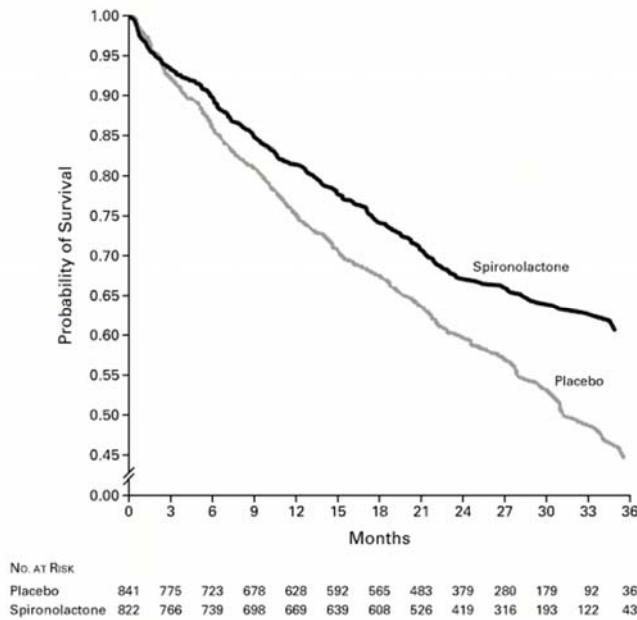


Box – plot interpretation

### Survival Curve

A survival curve is a special example of a line graph. It is sometimes referred to as a Kaplan-Meier Curve depending on the methodology used to construct the lines. Survival curves usually are used in studies of therapy, prevention, and prognosis. Survival per se need not be the dependent variable. For example, frequency of stroke, myocardial infarction, and tumor recurrence may be plotted on the y-axis of survival curves. Time is always on the x-axis (independent variable) of survival curves. It is not uncommon for there to be losses to follow-up in these studies owing because a patient may change address without notifying the investigators or withdraw consent for further

participation in the study or for any other reason. Because of losses to follow-up construction of a survival curve requires special methodology. Below is an example of a survival curve from a study comparing the use of spironolactone to placebo in patients with heart failure.



## Hypothesis Testing

In quantitative research, we usually ask if a relationship exists between variables. In order to answer that question, we propose a hypothesis and go about testing that hypothesis using an appropriate research design.

For example, we may wish to ask if there is a relationship between pneumococcal vaccination and the risk of developing community-acquired pneumonia. This type of pneumonia is typically caused by *Streptococcus pneumoniae* to which the vaccine is directed. We can hypothesize that those vaccinated are less likely than controls to get pneumonia. We may wish to ask adults from a population if they wish to participate in research and if so, indicate if they've received a pneumococcal vaccination. From these subjects, we randomly select 150 people who were vaccinated and 150 unvaccinated people. We can follow these groups for a pre-specified period and count cases of pneumonia. The results can be assembled in the following table.

	<b>Pneumonia</b>	<b>No pneumonia</b>
<b>Vaccinated</b>	10	140
<b>Not Vaccinated</b>	20	130

It would appear that our research hypothesis was correct. But, was this outcome due to the effect of the vaccination or were there other factors that could be responsible for this outcome? For example, were the vaccinated people just lucky this time? If we repeated this study several times, it is unlikely that we would get the exact same results every time. Indeed, the play of chance needs to be considered for most observational and interventional research. This is where statistical hypothesis testing comes in.

To analyze this data statistically, we use a standardized methodology called Hypothesis Testing. The first step in hypothesis testing is to propose a null hypothesis. The null hypothesis states that there is no difference between the groups with respect to some outcome. Put another way, the difference between the measures of an outcome between the groups is zero. The alternative hypothesis contradicts the null hypothesis and states that the difference between groups with respect to an outcome is not zero.

The next step is to statistically analyze the data using an appropriate methodology. For this data, a useful test is the Chi-squared test that can be used to analyze nominal data as we have in this dataset. This test is relatively easy to perform with pencil and paper or by using a web based calculator or statistical software.

The outcome of hypothesis testing is a 'p-value'. The p-value is the probability that the difference between the groups as large as or larger than the one observed in this dataset is due to the play of chance. The p-value is compared to a prespecified value, usually 0.05. The p-value for this dataset using the Chi-squared test was less than or equal to 0.10. That is to say, there was as much as a 10% chance that that the difference between the vaccinated and unvaccinated group's incidence of pneumonia was due to chance alone. Given that 0.10 is higher than the generally accepted level of significance of 0.05, we do not have enough evidence to reject the null hypothesis and we can indicate that the results of this study were not statistically significant.

Several statistical tests of hypothesis may be used to analyze data. The test that is appropriate to use depends on a number of factors including: the scale of measurement, the distribution of the data (parametric versus non-parametric data), the number of outcomes that have occurred (e.g. Fisher's exact test versus Chi-squared), and the complexity of the analysis (regression analysis) to name a few. In most cases, physicians will consult a biostatistician before embarking on a research protocol.

Some examples of statistical testing methods include: the Student's t-test, the paired t-test, the Chi-squared test, the Wilcoxon Rank Sum test, the Sign test, McNemar's test, log-rank test, and analysis of variance. Time permitting, these tests will be discussed in detail later in the course.

### **Type I and Type II Error**

Because we are dealing with probabilities, there is a chance that the results of hypothesis testing may be incorrect. Two types of error are recognized. A type I error is said to have occurred if the null hypothesis is rejected when in fact no difference in outcome exists between the groups. This is sometimes referred to as a rejection error or an alpha ( $\alpha$ ) error. The chance of making a type I error is equal to the level of significance that was specified prior to conducting the research. This value has been traditionally set at 0.05 or 5%. In other words, when we compare our p-value to a level of significance ( $\alpha$ ) of 0.05, we accept that the probability of committing a Type I error is as high as 1 chance in 20.

A type II error is said to have occurred if the null hypothesis is not rejected when in fact a difference in outcome between the groups exists. This is sometimes referred to as an error of missed opportunity because an association has been missed. Other terms for a type II error are acceptance error, or a beta ( $\beta$ ) error. We often focus on the value,  $1 - \beta$ . If there is a probability of making a type II error of 20%, then failing to reject the null hypothesis will be the correct course of action 80% of the time. The value,  $1 - \beta$  is known as the Power to detect a difference under the conditions of the trial.

You may have notice some quirked wording in this section. Hypothesis testing was designed to reject the null hypothesis or not reject the null hypothesis. It was not designed to accept the null hypothesis. There is a similar analogy in criminal trials where the defendant is found guilty (sufficient evidence to preclude reasonable doubt) or not guilty (insufficient evidence to preclude reasonable doubt. This method obviates the requirement to prove innocence before releasing the defendant.

### **Comparing Two Means from Independent Samples**

Most outcomes can be expressed with continuous data. Birth weight, blood pressure, serum creatinine are examples of continuous data. Means can be calculated for these outcomes, and when two independent groups are being compared, a t-test can be used for hypothesis testing.

For example, a researcher may wish to compare the serum ferritin levels between a sample of children with cystic fibrosis and a sample of healthy children. The null hypothesis would be that the serum iron in the cystic fibrosis group equals the serum iron in the healthy children. The blood is tested and data collected. A t-test will determine if any difference in serum iron between the groups may be accounted for by chance.

The t-test procedure varies slightly according to whether the variances of the outcome data are equal or unequal in the two samples. Statistical software will normally require some input regarding this issue.

### **Comparing Two Means from Paired Data**

Sometimes subjects may serve as their own controls. For example, in a study of an asthma drug, peak airflows can be measured before and after the use of a bronchodilator. The advantage of this method is that any statistically significant difference observed before and after medication use is likely due to medication effect.

The post peak flow can be subtracted from the pre peak flow. If the medication has no effect, the difference will be zero. The Paired t-test is the appropriate test statistic to use in this scenario. The null hypothesis in this case is, 'the difference between pre and post medication peak flows equals zero'.

### **Comparing Proportions**

Recall that it is not appropriate to calculate means from nominal data. Examples of nominal data may include outcomes that are binary (dead or alive), or non-binary (complications of surgery: death, myocardial infarction, deep vein thrombosis, hemorrhage, infection).

For these types of data, a chi-squared test can be used. The chi-squared test is relatively easy to perform and can be applied to tables that are larger than 2 by >2 in size (2 x k). For example, consider rates of infection from two surgical units.

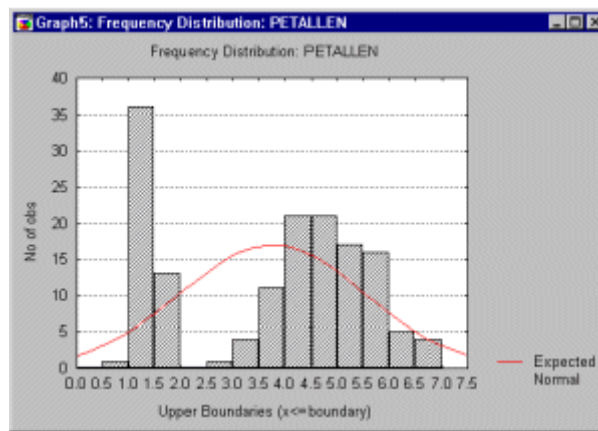
	<b>Pneumonia</b>	<b>Incision infection</b>	<b>Urinary tract infection</b>
<b>Unit A</b>	10	17	34
<b>Unit B</b>	19	20	20

The data was entered into a statistical program. The output revealed that the probability that a difference between the nursing units as large as or larger than the one observed here is due to chance is less than 0.05. In other words, the difference between the units is statistically significant.

### **Parametric and Nonparametric Tests of Statistical Hypothesis**

A distinction is made between data that is normally distributed (conforms to a bell-shaped curve) and data that is not normally distributed. This distinction is useful when we select statistical tests. Statistical tests that assume the data is normally distributed are classified as Parametric Tests. Examples of parametric tests include the t-test and Chi squared test.

When data is not normally distributed, nonparametric tests must be used. See the graph below that depicts data that is not normally distributed.



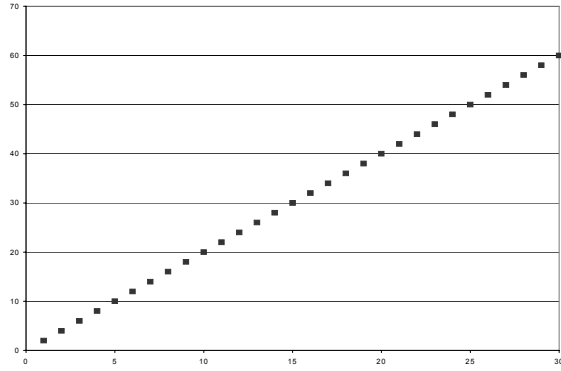
Examples of statistical tests used to analyze nonparametric data include the sign test and the Wilcoxon rank sum test.

When using nonparametric tests, no assumption need be made regarding the distribution of the data. Whereas users of parametric tests assume the data is normally distributed. A drawback to nonparametric tests is that they are less sensitive to statistically significant differences in the data.

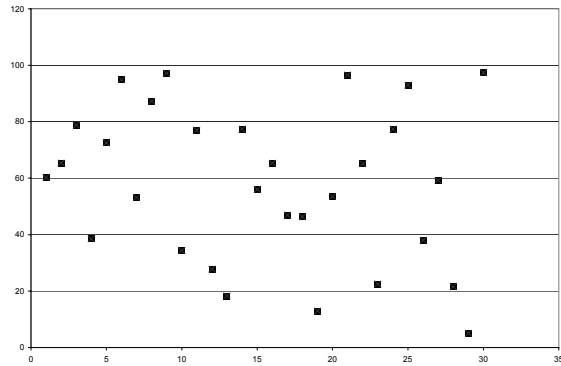
## Correlation

It is common to search for relationship between variables. One technique for measuring the strength and direction of an association between two variables is called correlation analysis.

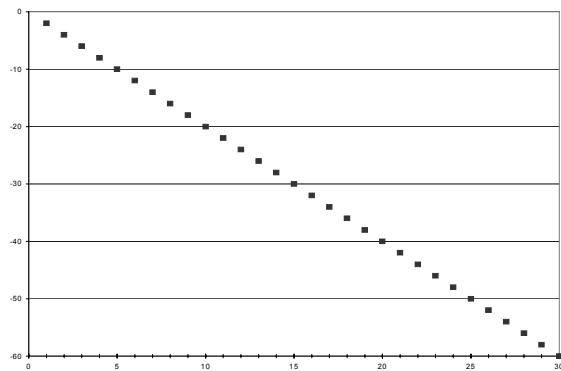
Although a number of statistical techniques are available, the Pearson's Correlation Coefficient is commonly reported. The outcome of this calculation ranges from  $-1$  to  $+1$ . A coefficient of  $+1$  expresses a perfect direct relationship between two variables;  $0$  expresses that no relationship exists between the variables and  $-1$  expresses that a perfect indirect relationship exists between the variables. Below are examples of how the data might appear under these circumstances.



Correlation = + 1



Correlation = 0



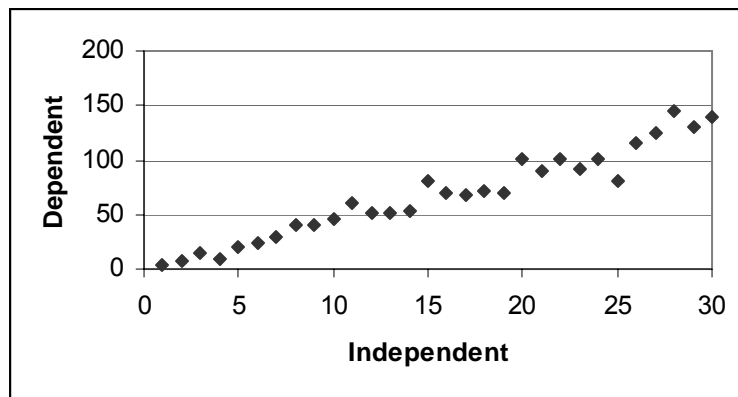
Correlation = -1

## Regression Analysis

It is increasingly common for investigators to analyze the relationship between variables linear regression and logistic regression. These techniques are sometimes referred to

as modeling.

The mathematics underlying these techniques is beyond the scope of the course. However, it is useful to know that regression analysis allow us to evaluate the effects of more than one independent (explanatory) variable on an outcome (response variable). For example, consider the graph below that shows the relationship between two variables.



Inspection of the data suggests that a direct (positive) relationship exists between the variables. However, the relationship is not perfect; the points do not form a line. By applying regression analysis to this data, an equation for the line that best fits the data can be calculated.

Linear regression is used when the outcome variable is continuous. Logistic regression is used when the outcome variable is non-continuous (e.g. presence or absence of stroke). The mathematics for regression analysis is complex and usually requires computer software designed for this purpose.

### **Sample Size Calculations**

When research is planned, consideration must be given to the number of subjects that should be included into the study. This is especially true when an intervention is being planned.

Consider a trial that evaluates a new antibiotic for the treatment of urinary tract infections. In this trial, patients with a urinary tract infection may be randomly allocated to receive the new antibiotic for 7 days or a standard antibiotic for 7 days. The outcome will be the proportion of patients cured. How many subjects would need to complete this trial if we wanted to obtain useful results? The results would be difficult to believe if only 2 people were enrolled. On the other hand, enrolling a million people seems excessive. A sample size calculation is needed.

Calculating the sample size was challenging prior to the advent of computers. Now, there is software available for this purpose. It is sufficient to know that with the input of

certain requirements and assumptions, a sample size may be obtained. Sample size calculations are available using internet-based software.

<http://calculators.stat.ucla.edu/powercalc/>

is an URL (internet address) you may wish to try or using a search engine, simply enter the terms Sample Size Calculator to find other sites.

The information required for calculating the sample size will vary somewhat according to the scale of measurement associated with the data. However, it is usual to specify a value set for alpha (the p-value considered to be significant, usually 0.05), beta (the acceptable chance of making a type II error, often 0.20), and the difference between groups that is felt to be important (or an estimate of the outcomes for each group). For a comparison of means, the anticipated standard deviation associated with the data is usually required.

Let's return to our urinary tract infection trial. It is known that the standard antibiotic is associated with a 75% cure rate and that it would be clinically meaningful if a new antibiotic could improve this cure rate by 15% to 90%. Let us also specify that we are willing to accept a Type I error rate of 5% and a Type II error rate of 20%. If we consult <http://calculators.stat.ucla.edu/powercalc/>, we see a number of choices. Because cure / not cure is an example of a binomial outcome and that we have two samples, we can use the binomial – two sample method.

The required data includes: our estimated outcome in the treatment (0.90) and control groups (0.75), level for alpha (0.05), and power (1-beta, 0.80). It also asks us to specify the number of tails in our experiment. This is two because we do not know with certainty that the new antibiotic will perform better than standard therapy.

After entering this information, we find that 97 subjects in each group would be needed to complete our trial in order to obtain results consistent with the levels of error that we specified. Given that some subjects may drop out of the trial early, say a 3% drop out rate, it is prudent to enroll 100 patients in each group.

There are advantages to selecting the right number of study subjects. If more subjects are enrolled than what is required to provide evidence of effect, then both exposure of subjects to risk of a new therapy and cost will be in excess of the ideal. If too few subjects are enrolled, the chance of type II error increases. Irrespective of any statistical consideration, physicians may view trials using small samples as not being representative of the population from which the sample was taken. For example, a trial that enrolls only 100 patients with diabetes mellitus is not likely to represent patients with diabetes in general owing to the wide variation in the characteristics associated with patients with this condition.

## **UNIT 3: RESEARCH DESIGN STUDY NOTES**

The purpose of the research design unit is for the student to survey the common designs used in clinical research. Advantages and disadvantages of selected research designs will be stressed. The roles of bias, confounding, and chance will be covered.

Research design refers to the 'architecture' or 'what is done' during the course of the research project. Clinicians (physicians that see patients) need to be knowledgeable about research design in order to evaluate the validity, results, and applicability of medical research to their practices. Research may be classified along selected characteristics.

### **Quantitative Research**

In a broad sense, quantitative research refers to investigations where dependent and independent variables may be expressed numerically. Degree of hypertension, frequency of infection, birth weight, presence or absence of smoking can be expressed numerically. This compares with qualitative research where the variables under consideration are better described with words. For example, attitudes toward illness, coping strategies used to deal with loss, or gaps in communication are better described with words than numbers.

How one goes about answering a question is a matter of research methodology. In medicine, there are a number of research designs that are commonly used in both the quantitative and qualitative research traditions. We shall first describe issues relevant to quantitative research.

### **Observational versus Interventional Research**

Within the quantitative research tradition, a distinction is made between observational and intervention studies. Observational studies are those where the investigator observes but does not control the variables (particularly independent variables) under consideration. Correlational studies, case reports, case-series, cross-sectional surveys, case-control, and cohort studies (retrospective / prospective) are classified as observational studies.

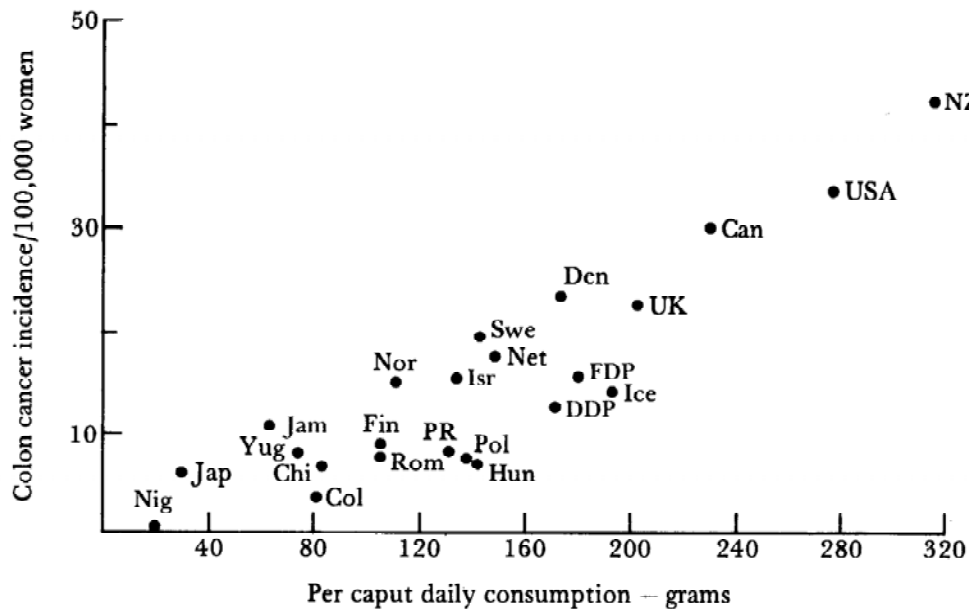
Intervention trials (clinical trials) are those where the investigator determines (to some extent) an exposure under consideration. The term 'exposure' is used in the broadest sense to include any variable that might influence an outcome under consideration.

To illustrate the distinction between observational and intervention research consider the evaluation of a vaccine as measured by the association between being vaccinated and protection against infection. In an observational trial, the researcher would simply measure and compare the rates of vaccination and infection in two or more populations (or the same population at different times). In an intervention trial, an investigator would

allocate people to either receive a vaccination or not receive a vaccination. Rates of infection would then be compared between the vaccinated and unvaccinated people.

### Correlational Study

In a correlational study (sometimes called ecologic study), data from entire populations are used to compare disease frequencies between different groups during the same period of time or in the same population at different points in time. For example, shown below is the incidence of colon cancer from several countries plotted against the per capita dietary meat intake.



There appears to be a correlation between colon cancer incidence and per capita daily meat consumption.

There are two significant advantages to this research design: 1) it is inexpensive and easy to obtain this data and 2) the research poses no risk to the population under study.

There are disadvantages as well. It is not known if the people who developed colon cancer actually ate more meat. We only have population averages. Another disadvantage is that we cannot ascertain an association is causal from a correlational study. For example, a correlational study is likely to find that diabetes mellitus is more frequent among countries that have a high prevalence of obesity. However, from this data is not possible to know is obesity causes diabetes mellitus or if diabetes mellitus causes obesity.

These disadvantages pose significant threats to validity. Therefore, correlational

research is often viewed as hypothesis generating (raises a possibility) rather than hypothesis confirming.

### **Case Reports and Case Series**

The case report is the most basic type of descriptive study of individuals. It consists of a detailed report surrounding the course of a single patient. An example of a case report is the description of a female who developed a pulmonary embolism 5 days after an estrogen containing preparation was initiated to treat endometriosis. The author questioned the possibility that the adverse outcome was associated with the intervention. This case report was published in 1961. Research established this association many years later.

A case report can be expanded into a case series when more than one patient with similar features are described. For example, a case series described 5 young healthy men that had developed *Pneumocystis carinii* pneumonia in the Los Angeles area during a 6-month period in 1980-1981. This was an unusual clustering of cases. Subsequently, it was determined that these patients were suffering from what was to be known as the Acquired Immunodeficiency Syndrome.

The advantages of case reports and case series include low cost, no risk to the subject (no intervention occurs for research purposes), and completion of the study over a relatively short duration. The disadvantage of this form of observational research is that the cases seldom represent any pre-defined population and as such, it is impossible to know the frequency of observed events. There is no denominator from which to define the frequency of the reported cases.

### **Cross-Sectional Surveys**

The objective of a cross-sectional survey is to simultaneously classify subjects according to disease status and exposure status. Data may be collected in-person or by any medium of communication (telephone, letter, internet based).

For example, body measurements and fasting blood glucose are taken at the same encounter. From this data, the presence of obesity and diabetes may be determined and an association between risk factor and disease can be calculated.

The advantage of this study design includes low cost, minimal risk to subjects, and short period to complete the research. The main disadvantage is that there is no temporal sequence in the data. In our example, it is no more possible to conclude that obesity causes diabetes than to conclude that diabetes causes obesity. In fact, it is common for people with diabetes mellitus to gain weight with treatment. You may have noticed that cross-sectional surveys consider the individual and correlational study consider populations.

The data from a cross-sectional survey may be placed in a two by two table. Consider

hypothetical data from our obesity.

	<b>not high blood glucose</b>	<b>high blood glucose</b>
<b>not obese</b>	99	1
<b>obese</b>	85	15

From this data, we can calculate the strength of the association between obesity and high blood glucose using the methods described in the biostatistics section.

### **Case-control Study**

In a case-control study, a group of individuals with a disease are identified and compared to a control population with respect to the presence or absence of an exposure.

An example of this research design would include a published study of the relationship between stroke and the quality of hypertension control in England. In this study, patients who had a stroke were identified from a stroke register that was maintained by the local health authority. For each stroke patient, two patients were selected at random from the community who had the same age and sex. As an aside, these controls are referred to as 'age and sex matched controls.' Then, the presence or absence of a history of hypertension was determined from the patients' health records.

It was found that in 61% of the 267 cases of stroke, there was a history of hypertension. In 42% of the 534 controls, there was a history of hypertension. A history of hypertension was more frequent among the stroke patients than controls.

Below is the case-control method.

CASES (Disease +) -----→ ? frequency of risk factor

CONTROLS (Disease -) -----→ ? frequency of risk factor

Notice that in a case-control study, we start with cases / controls and then determine if a risk factor was present by working back in time from previous records. This is known as a retrospective study because the data used was actually collected prior to initiating the research.

The advantages of case-control studies include:

- 1) a relatively low cost and short duration required to complete the study,
- 2) it is suited to the evaluation of diseases with long latent (clinically unapparent) periods,
- 3) it is optimal for the evaluation of rare diseases,

- 4) it can examine multiple exposures for a single disease, and
- 5) does not impart a risk to the patients.

Disadvantages of case-control studies include:

- 1) disease incidence cannot be measured from a case-control studies unless the study is population based which is seldom possible,
- 2) it may be hard to establish a temporal relationship between exposure and disease in some cases,
- 3) it is particularly prone to selection and recall bias, and
- 4) it is not efficient for the evaluation of rare exposures.

### **Cohort Study (Retrospective and Prospective)**

In a cohort studies, groups of people who are exposed to a given risk factor are compared to people who are unexposed (controls) to that risk factor with respect to the incidence of disease. This may be done in a retrospective or prospective manner depending according to the relationship between data collection and the initiation of the study.

Consider a study where the relationship between coal mining and lung disease is under investigation. In a retrospective cohort study, this relationship is assessed by identifying people who worked in a coal mine 25 years ago (exposed) from company records and identifying people who lived in the same town 25 years ago but worked in places other than a coal mine (non-exposed or controls). Then, the health records of both cohorts of people can be reviewed for the development of lung disease. The incidence of lung disease in each group is calculated and compared.

To perform this study prospectively, the investigators would identify people currently working in a coalmine (exposed) and identify those who currently work in places other than a coalmine. Then, investigators follow both groups of people for 25 years while monitoring for new cases of lung disease. The incidence of lung disease in each group is calculated and compared.

Below is the cohort method.

RISK FACTOR (Exposed) -----→ ? Incidence of Disease  
 NO RISK FACTOR (Non-exposed) -----→ ? Incidence of Disease

Notice that in a cohort study, we start with exposed and non-exposed populations and determine the incidence of disease thereafter.



of about 1 million people. However, it is feasible to sample the population of persons with diabetes mellitus. Sampling may be done in a variety of ways. Subjects may be approached at random to participate in a clinical trial (random selection). Subjects could also respond to advertisements or to being approached in a variety of clinical settings.

It is important that the population under consideration be defined prior to initiating the trial. This is done by specifying inclusion and/or exclusion criteria. For example, in a trial that evaluates a medication for rheumatoid arthritis, it would be appropriate to specify that subjects meet all of the following entry:

- 1) have rheumatoid arthritis as defined by standard criteria,
- 2) have disease that is currently not under control,
- 3) if female, not pregnant and/or using birth control as appropriate, and
- 4) are free from other conditions that increase the risk of the study medication.

### Subject Allocation

There are different methods for allocating subjects to the available study arms in a trial. Patients may be allocated to one study arm according to subject or investigator preference. Alternatively, subjects may be allocated at random according to a coin toss, random number table, computerized random number generator.

Random allocation is a powerful technique! It cannot be overemphasized that random allocation is the best tool we have to distribute both the KNOWN and UNKNOWN determinants of disease equally among study arms.

Random allocation does not restrict the study arms to having equal numbers of subjects. For example, it is possible to randomly allocate 2 subjects into a study arm for every 1 subject allocated to a control group.

The method of random allocation can vary. Although coin toss is conceptually appealing, in practice it is rarely used to allocate subjects. A coin toss may not produce the desired results particularly when the study sample is small. Consider a trial of 40 subjects (30 male and 10 female) who are to be randomly allocated to one of two study arms. If the outcome is influenced by gender, it would be desirable that the ratio of male and female in both groups be equal. A method for ensuring equal gender proportions would be to label 40 identical cards as follows: 15 cards – male/control, 15 cards male/experimental, 5 cards female/control, and 5 cards female/experimental. The male cards are shuffled and placed in one box; the female cards are shuffled and placed in a second box. When a male is enrolled, a card is pulled from the male box and similarly for a female enrollee.

### Levels of Blindedness (Concealment, Masking)

Blinding (concealment, masking) refers to preventing subjects and/or investigators from knowing the allocation of the study subjects. There are different levels of blinding. In an open-label study, the study arm to which a subject is allocated is known by subjects and investigators. In a single-blind trial, the subjects are prevented from knowing which study arm they have been allocated to. This is accomplished by using a placebo (identical appearing but inert medication) or sham procedure (identical appearing but ineffective procedure) in the control arm of the trial. Sometimes, blinding procedures can be very elaborate.

In a double-blind trial, neither the investigators who are measuring the outcomes nor the subjects have knowledge of allocation.

In a triple-blind trial, subjects, investigators measuring outcomes, and those analyzing the data are unaware of subject allocation until an allocation code is broken once the trial is completed and the data analyzed. In double and triple-blinded trials, a person not involved in the trial is trusted to maintain the allocation code. The rationale for blinding is to reduce bias. This is discussed under Critical Appraisal of Articles about Therapy.

### Quasi-experimental Designs

For completeness, it should be mentioned that there exists a series of interventional study designs known as quasi-experimental designs. Quasi-experimental designs are several and vary considerably. A feature common to many quasi-experimental study designs is that allocation is based on some subject characteristic rather than by play of chance as in the case of a randomized controlled trial.

An example of a quasi-experimental design is the non-equivalent groups design. In this design, subjects with the same condition are selected according to some characteristic such as the hospital they are admitted to. The severity of their disease is measured and an intervention is applied to those at one hospital only. Change in disease severity is measured later and compared. The problem with this method is that factors other than the intervention (equipment, expertise) may be different at the two hospitals.

Below is a summary of the non-equivalent groups design.

Group A → Measure Disease → Intervention → Measure Disease

Group B → Measure Disease → No Intervention → Measure Disease

### **Common Threats to the Research Validity and Reliability**

In this section, threats to validity and reliability will be discussed. Although these concepts have primarily been developed in the context of quantitative research, some features may be applicable to qualitative research.

### Inference

In research, we must commonly draw conclusions about a population based on the evidence available. It is virtually impossible to study an entire population of individuals with a given disease or exposure. Therefore, we must make inferences about populations by collecting and studying samples of those with the disease or exposure of interest. Researchers assume that there will be a difference between results obtained from a sample and results obtained if an entire population were studied. These differences can be accounted for by one or more of the effects of bias, confounding, or the play of chance. To understand these factors, it is worthwhile to review the concepts of validity and reliability.

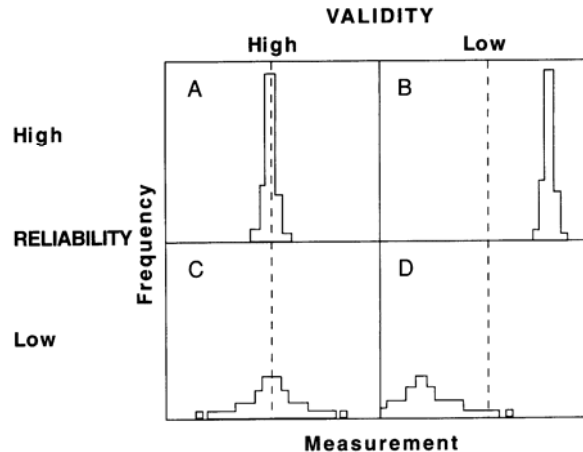
### Validity and Reliability

Validity and reliability refer to the ability to measure variables. Consider a trial where a new procedure for measuring serum sodium is being assessed. Assume that the mean serum sodium for people living in Calgary is 145 mmol/l and that several samples of 100 Calgarians was selected at random from which blood was drawn and analyzed using the new procedure.

The validity of a measure is the extent to which differences in the results of the measurement reflect true differences among individuals on the characteristic that is supposed to be measured. If the mean value for serum sodium in our samples was 145 mmol/l, we can judge that the new procedure is valid. If the average values obtained from our samples were 130 mmol/l or 160 mmol/l, we would judge the test invalid.

The reliability of a measure is its consistency. If the mean value of serum sodium for each sample of Calgarians ranged between 142 and 144 mmol/l, we would judge the new test to be reliable even if it was not valid. On the other hand, if the mean value from each sample ranges from 130 to 160 mmol/l, we would conclude that the new procedure was unreliable to some extent even if the average value obtained from all the samples was 145 mmol/l.

In the diagram below, we see examples of tests that have high and low validity and reliability. The two characteristics are independent. A test can have high validity and low reliability and so forth.



## Bias

Bias is a process at any stage of research that produces results that depart systematically from the true values. Bias tends to threaten validity. The classification of bias varies among authors (this is unfortunate) and many forms of bias have been described.

Selection bias occurs when comparisons are made between groups of patients that differ in determinants of outcome other than the one under study. For example, a cohort trial might include an exposed older population from one town and a non-exposed younger population from another town. Rates of degenerative disease will be more likely in the exposed population owing to the bias imparted by age irrespective of any risk attributable to the exposure.

Measurement bias occurs when the methods of measurement are dissimilar among groups of patients. This might occur when historical cohorts are used for comparison groups. For example, one might wish to compare rates of brain tumor before and after the introduction of cellular telephone technology. Given that CT scanning which is a sensitive test for detecting brain tumors was not widely available prior to the invention of the cellular phone, this study would be vulnerable to measurement bias. Recall bias is similar to measurement bias. Subjects are more likely to recall recent events than distant events and subjects are often more likely to associate a recent exposure to disease than a distant exposure. Misclassification bias refers to errors in classifying subjects by either disease or exposure status. Measurement or recall bias could cause misclassification.

It is important to understand how bias may affect the results of an investigation. It is common to express the direction of bias as being toward or away from the null hypothesis. The null hypothesis states that no difference exists between the groups under consideration. In our cell phone and brain tumor example, the measurement bias

associated with lack of readily available CT scanning will cause an underestimation of the incidence of brain tumor during the time prior to the introduction of cellular telephones. This will bias the result away from the null hypothesis that there is no difference in the incidence of brain tumors within the population prior to and after the introduction of cellular phones.

### Intention to Treat Analysis versus Explanatory Analysis

Bias may occur during the analysis phase of research. How data is analyzed may impact on the results obtained. In clinical trials, it is usual for investigators to describe how they handled data arising from subjects who did not complete the treatment to which they were allocated. For example, think of a randomized clinical trial that compares the effects of chemotherapy and surgery among patients with breast cancer. In such a trial is likely that that a proportion of subjects allocated to surgery (say 15%) will have undergone surgery at the trial's conclusion. The usual reason for this is deterioration in the subjects' health status between the time of allocation and the date of surgery. Fearing that the sicker patients may not survive surgery, the surgeon or anesthesiologist will cancel surgery. Data from this study can be analyzed in two ways.

In the Intention to Treat Analysis, patients will be analyzed according to the group to which they were allocated. In the Explanatory Analysis, patient will be analyzed according to intervention they actually received. Which is better?

Most investigators use the Intention to Treat Analysis. In our example, there are two risks associated with allocation to surgery. First, there was a delay between allocation and the date of surgery. Second, surgery cannot be undertaken if the patients do not meet some test of wellness. This is usually less true for chemotherapy that may be more available and less dependent on patient health status. If we were to exclude the patients who did not receive surgery, we would unfairly bias in favor of surgical therapy.

### Publication Bias

Bias may also occur after the trial has been completed. Publication bias refers to the tendency of investigators and publications to report the results of studies where a positive result was obtained in favor of studies where a negative result was obtained. Negative results, in this context, refer to studies where the results were not statistically significant. Publication bias is difficult to prove because it is not usually possible to know of all trials that have failed to be published owing to a negative outcome.

Hence, the presence of publication bias is usually a matter of controversy among the experts in the field. By way of example, in a recent Canadian Journal of Emergency Medicine (CJEM 2001;3(3)) the authors suggest that proponents of a clot-dissolving drug used in the setting of stroke may have been vulnerable to publication bias.

### Confounding

Confounding occurs when two factors are associated and the effect of one is confused with or distorted by the effect of the other. Some authors classify confounding as a form of bias although most consider confounding to be a special form of bias.

For confounding to be present, two exposures must be associated with each other but only one of the exposures is a risk factor for disease. A classic example of confounding is the relationship between alcohol use and lung cancer. If one did a case-control study comparing the rates of alcohol consumption among those with and without lung cancer, it is likely that an association will be found between this exposure and disease.

However, if one stratifies these populations by the presence or absence of a history of cigarette use, we will find that among cigarette users, the relationship between alcohol use and lung cancer is strong and among non-smokers the relationship between alcohol use and lung cancer is not present. The explanation for these findings is confounding. It turns out that those who consume alcohol smoke more than those who do not use alcohol. Hence, our original finding of a relationship between alcohol consumption and lung cancer was incorrect.

### Role of Chance

The play of chance may be responsible for sample results that differ from that of the underlying population. This will be covered in some detail in the biostatistics section. For this section, it is sufficient to point out that the role of chance can be reduced by increasing the size of the sample taken from the population. This is both intuitive (make sense) and is demonstrable mathematically. Recall that variance and standard deviation get smaller when the sample size increases.

As indicated earlier, random allocation is a powerful tool because it attempts to distribute the known and unknown determinants of disease equally among the study arms in a clinical trial. However, it should be pointed out that the chance that random allocation will be successful increases when the sample size increases.

## **Qualitative Research**

### What is Qualitative Research?

Qualitative Research is a systematic process, based on distinct methodological traditions of inquiry that explores and seeks understanding of social or human problems. A qualitative researcher analyzes words to build a complex, holistic picture, reports detailed views of informants, and conducts the study in a natural setting.

Qualitative research methods were developed in the social sciences to enable researchers to study social and cultural phenomena. Examples of qualitative research traditions include: grounded theory, case studies, ethnography and phenomenology. Qualitative data sources include observation and participant observation (fieldwork), interviews and questionnaires, documents and texts, and the researcher's impressions

and reactions.

Qualitative research builds on the one thing that distinguishes humans in the natural world, our ability to talk! Qualitative research methods are designed to help researchers understand people and the social and cultural contexts within which they live. Kaplan and Maxwell argue that the goal of understanding a phenomenon from the point of view of the participants as well as particular social and institutional context is largely lost when textual data are quantified.

Purposes of Qualitative Research can be summarized as: understanding the meaning, understanding the context, identifying the unanticipated, and developing theory.

### Contrasting Quantitative and Qualitative Research

The following table lists the attributes of quantitative and qualitative research.

<b>Domains</b>	<b>Qualitative</b>	<b>Quantitative</b>
Ontology	Multiple realities	Single objective reality
Epistemology	Knowledge based upon individual perception and understanding	Knowledge as a part of reality that is separate and independent from individuals
Type of Reasoning	Inductive	Deductive
Purpose	Reveal complexity Uncover meanings of human experience Theory generating	Predict Explain  Theory testing
Type of Question	Open-ended Process oriented	Pre-specified Outcome oriented
Type of Analysis	Narrative description Constant comparison	Numerical estimation Statistical inference
Context	Natural setting	Controlled setting
Relationship of the Researcher	Researcher interacts directly with those they study	Researcher remains distant and independent, neutral
Language of verification	Credibility Transferability Dependability Confirmability	Internal Validity External Validity Reliability Objectivity

**Table 1. Differentiating Qualitative and Quantitative Research**

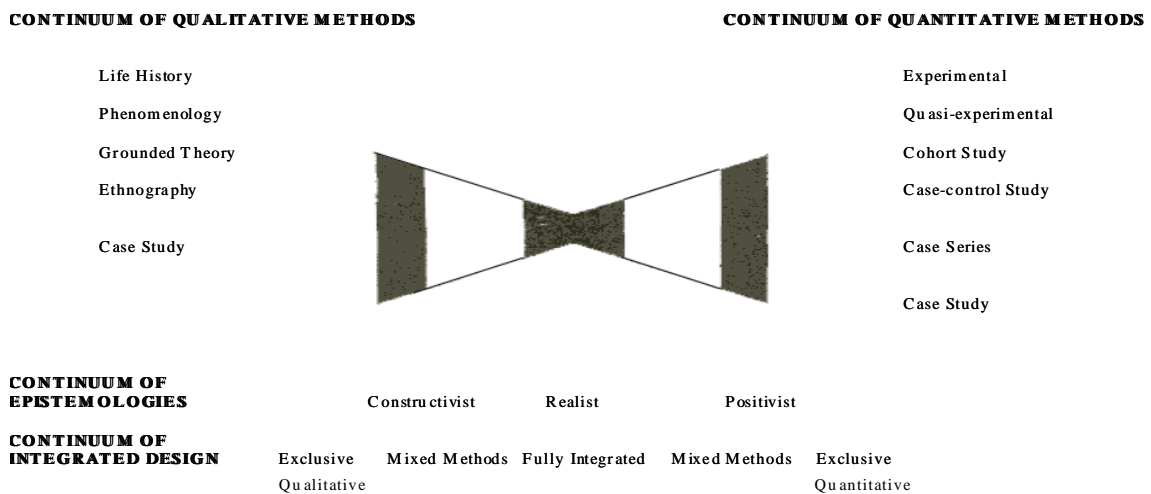
Adapted from: Casebeer & Verhoef, 1997; Creswell, 1994; DePoy & Gitlin, 1994; Krefting, 1991.

### Assumptions Underlying Qualitative Research

All research (whether quantitative or qualitative) is based on some underlying assumptions about what constitutes 'valid' research and which research methods are

appropriate. In order to conduct and/or evaluate qualitative research, it is therefore important to know what these (sometimes hidden) assumptions are. There are three common philosophies associated with qualitative research. Each philosophy has associated assumptions.

**Positivist Assumptions:** Positivists generally assume that reality is objectively given and can be described by measurable properties that are independent of the observer (researcher) and his or her instruments. Positivist studies generally attempt to test theory, in an attempt to increase the predictive understanding of phenomena. Most quantitative research designs use a positivist approach.



**Figure 1. The research continua**  
Adapted from: DePoy & Gitlin, 1994, p. 23

**Interpretive Assumptions:** Interpretive researchers start out with the assumption that understanding what reality means to someone else can only be achieved through social constructions such as language, consciousness and shared meanings.

Interpretive studies generally attempt to understand phenomena through the meanings that people assign to them. Interpretive research does not predefine dependent and independent variables, but focuses on how humans make sense of what is happening as the situation emerges. Many qualitative traditions use an interpretive approach

### Qualitative Research Methods

Just as there are various philosophical perspectives that can inform qualitative research, so there are various qualitative research traditions. A research tradition is a strategy of inquiry which moves from the underlying philosophical assumptions to

research design and data collection. The choice of research method influences the way in which the researcher collects data. Specific research methods also imply different skills, assumptions, and research practices. The four research methods that will be discussed here are, case study research, ethnography, and grounded theory.

**Grounded Theory:** Grounded theory is a research method that seeks to develop theory that is grounded in data systematically gathered and analyzed. According to Martin and Turner, grounded theory is "an inductive, theory discovery methodology that allows the researcher to develop a theoretical account of the general features of a topic while simultaneously grounding the account in empirical observations or data." The major difference between grounded theory and other methods is its specific approach to theory development - grounded theory suggests that there should be a continuous interplay between data collection and analysis.

**Case Study Research:** The term "case study" has multiple meanings. It can be used to describe a unit of analysis (e.g. a case study of a particular organization) or to describe a research method. The discussion here concerns the use of the case study as a research method. Although there are numerous definitions, Yin defines the scope of a case study as follows: "A case study is an empirical inquiry that: investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident".

**Ethnography:** Ethnographic research comes from the discipline of social and cultural anthropology where an ethnographer is required to spend a significant amount of time in the field. Ethnographers immerse themselves in the lives of the people they study and seek to place the phenomena studied in their social and cultural context.

**Phenomenology:** The central tenet of phenomenology is to determine what an experience means for the persons who have had the experience and are able to provide a comprehensive description of it. From the individual descriptions, general or universal meanings are derived, in other words, the essence of structures of the experience

### Qualitative Techniques for Data Collection

Each of the research methods discussed above uses one or more techniques for collecting empirical information or data. These techniques range from interviews, focus groups, observational techniques such as participant observation and fieldwork, through to archival research. Written data sources can include published and unpublished documents, company reports, memos, letters, reports, email messages, faxes, newspaper articles, and so forth.

### Sampling Strategies in Qualitative Research

**Purposive Sampling:** select information-rich cases for in-depth study to examine meanings, interpretations, processes, and theory. Examples include:

Maximum Variation where wide variations in the experience or process being examined is desirable,

Typical Case Sampling which may be used if the unit of analysis is large e.g. a village,

Criterion Sampling where all cases meet a set of criterion e.g. everyone discharged from an acute care hospital who received short-term home care during a selected week, and

Stratified Purposive Sampling where one is not trying to generalize to a population but rather there is an attempt to stratify on the basis of dimensions that might be important to the phenomenon being studied e.g. living with HIV / AIDS, include both men and women

### Analyzing Qualitative Data

Although a clear distinction between data gathering and data analysis is commonly made in quantitative research, such a distinction is problematic for many qualitative researchers. Therefore it is perhaps more accurate to speak of "modes of analysis" rather than "data analysis" in qualitative research. These modes of analysis are different approaches to gathering, analyzing and interpreting qualitative data. The common thread is that all qualitative modes of analysis are concerned primarily with analyzing words or text (whether verbal or written).

### Maintaining Rigor in Qualitative Research

It is important to maintain rigor in qualitative research. There are four important considerations relevant to qualitative research.

Neutrality (Objectivity): This refers to freedom from researcher biases,

Dependability (Reliability): This refers to consistency of the data gathering and analysis,

Credibility (Internal Validity): This asks if the findings describe an accurate portrait,

Transferability (Generalizability): This refers to the applicability of the conclusions.

### Conclusion

Qualitative research must respect the principles of all good research. It must serve a useful purpose supported by relevant literature, use an approach that is appropriate for the subject and the state of the knowledge about the topic under investigation with an appropriate method and attend to maintaining the quality or rigor of the research.

## **UNIT 5: EVIDENCE BASED MEDICINE STUDY NOTES**

The purpose of the evidence based medicine unit is for the student to learn how to formulate a researchable question from a clinical scenario, effectively search for and critically appraise evidence and apply what has been learned. This unit builds on the knowledge and skills learned in units 1 through 4 and require the student to complete a Critically Appraised Topic.

### **What is Evidence Based Medicine?**

Evidence Based Medicine (EBM) is about solving clinical problems. Specifically, EBM is the gathering and application of the highest quality scientific evidence to clinical problems in the context of the values and judgments of those affected by clinical problems. EBM is a relatively new phenomenon and it may be surprising that the term, 'Evidence Based Medicine' was introduced around 1990.

There are specific steps that form the process of EBM.

- 1) From a patient encounter, a researchable clinical question is posed.
- 2) Evidence addressing the clinical question is gathered using a number of information technologies.
- 3) The evidence gathered is critically appraised which includes an assessment of the validity, results, and applicability.
- 4) The highest quality evidence is applied to the patient encounter in the context of the patient's values and physician's judgments regarding appropriateness.

Some may add two additional steps:

- 5) What was learned is stored for future access.
- 6) The outcomes associated with any decision are assessed.

There are two fundamental principles of EBM. First, evidence must be applied in the context of the values and judgments of those affected by the clinical problem. For example, there may be irrefutable evidence supporting the benefit of antibiotics in the setting of pneumonia. However, in the case of a dying patient who wishes to receive only palliative care, it may be entirely appropriate to not use antibiotics should pneumonia develop. The second principle is that there exists a hierarchy of medical evidence. Any observation about the apparent relationship between variables may constitute potential medical evidence. We may make observations in the context of several research designs. Each design is associated with advantages and disadvantages that influence the validity, results, and applicability of the research and together impact on the overall quality of the evidence.

## **Formulating a Researchable Clinical Question**

It is common and desirable to use actual or hypothetical clinical scenarios to improve on the applicability of what is learned in medicine. Consider the following clinical scenario.

A 55-year-old male was encouraged to see a physician after his blood pressure was checked at the local shopping mall. He feels well and has no history of cardiovascular disease. He is a non-smoker, takes no medications, and reports no allergies. He is found to have a blood pressure of 170 / 110 mmHg and subsequent visits reveal similar readings. You make a diagnosis of hypertension (high blood pressure). He wonders if he should be taking medication to prevent stroke and heart attack.

A number of clinical questions can arise from a scenario. It is usual to consider the patient's concern(s) first. In this scenario, the patient wonders about taking blood pressure medication. It is important to make the question as specific as possible to avoid confusion and enhance the ability to find answers. In this scenario, a specific researchable clinical question might be, "in a 55 year old male, what are the effects of treating hypertension on the incidence of cardiovascular disease in comparison to no treatment?" Other lines of inquiry could include, the role of non-pharmacological (lifestyle) interventions, the effects of different classes of anti-hypertensive medications, and investigations that are appropriate for the hypertensive patient.

A well-structured clinical question usually contains three components:

- 1) the population of interest,
- 2) the intervention and comparator, and
- 3) the outcome.

It is also important to classify the clinical question as one of harm, diagnosis, therapy / prevention, or prognosis. This is done because the criteria used to appraise the evidence vary with this classification scheme. Other questions may be considered including an assessment of costs associated with an intervention.

## **Searching for Evidence**

Once a researchable clinical question is formulated, the search for answers begins. Medical information can be collected from many sources including the lay press, peers, expert opinion, textbooks, primary research articles, systematic reviews, and clinical practice guidelines. This list is not exhaustive.

A principle of Evidence Based Medicine is that we primarily consider best quality evidence. A number of medical organizations (American Thoracic Society, Canadian Diabetes Association) have developed tools for assessing the quality of the medical evidence. Overwhelmingly, these tools (rating scales) consider research design to be

the most important indicator of evidence quality. If we consider questions of therapy or prevention, recommendations that are based on randomized controlled clinical trials offer the highest quality of evidence, recommendations based on observational research is of intermediate quality, and recommendations based on case series, case reports or expert opinion in the absence of supporting clinical research offers the lowest quality of medical evidence. Given this classification scheme, it is possible to identify and compare the quality of available medical evidence.

It is increasingly common for physicians to use on-line sources of medical information. These sources may be 'pre-filtered' or 'unfiltered'. Pre-filtered sources are those where the authors have gathered and summarized the information required. On-line textbooks and information services such as Up-to-Date are examples. Unfiltered sources of medical information include the U.S. National Library of Medicine's MEDLINE database. This database offers access to over 11 million citations of clinical and pre-clinical research.

The decision to use an information source depends on the objective of the reader. During the early stages of medical training, students need to learn about human physiology, anatomy, and pathophysiology. Textbooks and pre-filtered sources tend to be more rewarding. As medical training proceeds, the clinical questions become more specific and unfiltered sources become more important.

The science of searching for medical information is known as Clinical Informatics.

### **Introduction to Critical Appraisal**

In the paradigm of EBM, evidence gathered should be critically appraised. Critical Appraisal refers to an assessment of validity, the results obtained, and the applicability of the evidence. A series of 'User's Guides to the Medical Literature' has been developed which provide a template by which medical evidence is critically appraised. Each type of clinical question has a corresponding template.

Common to the templates is a focus on:

- 1) the validity of the research,
- 2) the results of the research, and
- 3) the applicability of the results to the clinical problem.

The User's Guides were developed primarily by Canadian Physicians. They have been published widely and may be accessed at the Centre for Health Evidence website.

### **Critical Appraisal – Articles about Therapy or Prevention**

In assessing an article about therapy or prevention, three areas are considered.

Are the results valid?

Validity refers depends on whether the research was designed in a way that justifies the claims about the benefits or risks of a therapeutic or preventative regimen. The following questions are asked to assess validity.

- 1) Was the assignment of patients to study arms randomized? Randomization offers the best chance that factors KNOWN and UNKNOWN to influence the outcome under interest have been distributed equally among study arms.
- 2) Were all patients who entered the trial properly accounted for and attributed at its conclusion? Losses to follow-up may bias the results.
- 3) Were patients analyzed in the groups to which they were randomized? This refers to analysis according to intention to treat.
- 4) Were patients, health workers, and study personnel "blind" to treatment? Concealment or blinding reduces biases that the subjects, researchers, or analysts may bring to the study.
- 5) Were the groups similar at the start of the trial? Typical the 'Table 1' assessment of subject attributes addresses this.
- 6) Aside from the experimental intervention, were the groups treated equally?

#### What are the results?

It is key to distinguish between clinically significant and statistically significant effects of an intervention. Clinical significance implies that the benefits of an intervention are worth any associated risk or cost. Statistical significance refers to the probability making a type I statistical error. For an intervention to be considered beneficial (or harmful) both clinical and statistical significance is required.

- 1) How large was the treatment effect? This question asks about the point estimate associated with an intervention. Recall that treatment effects may be expressed in several ways (relative risk, absolute risk reduction, number needed to treat, odds ratios). Larger treatment effects provide better evidence of efficacy than smaller treatment effects.
- 2) How precise was the estimate of the treatment effect? The confidence interval expresses the precision associated with the point estimate. Narrow confidence intervals provide better precision than wide confidence intervals. Confidence intervals that cover an odds ratio of 1.0, a relative risk of 1.0, or a difference of 0.0 imply that no statistically significant difference between experimental (active treatment) and control (placebo) groups.

#### How can I apply the results to patient care?

Applicability of the evidence is important. Three questions assess applicability.

- 1) Were the study patients similar to my patient? This question refers to the generalizability of the study to the patient under consideration.
- 2) Were all clinically important outcomes considered? It is usual for published research to emphasize the primary outcome such as a treatment efficacy. However, frequency of adverse effects of an intervention or cost may receive less attention. These outcomes may be just as or more important to patients.
- 3) Are the likely treatment benefits worth the potential harm and costs? This question compares treatment efficacy and risk of adverse effect and cost. Number needed to treat can be linked to number needed to harm.

### **Critical Appraisal – Articles about Diagnostic Tests**

In assessing an article about a diagnostic test, three areas are considered.

#### Are the results valid?

- 1) Did clinicians face diagnostic uncertainty? It is easy to tell the difference between a perfectly healthy patient and one severely affected by a disease. However, the utility of a test lays in its ability to identify disease states among patients with similar presentations. In this context, diagnostic uncertainty is likely to be high. For example, a test that distinguishes between the chest pain of coronary disease and the chest pain of esophageal spasm would be very useful.
- 2) Was there a blind comparison with an independent gold standard applied similarly to the treatment group and to the control group? It is key that in a study of a diagnostic test that a gold standard is applied to the subjects receiving the experimental and standard diagnostic protocols. In the absence of the application of a gold standard, it is not possible to be certain that all patients have been appropriately classified with respect to disease status.
- 3) Did the results of the test being evaluated influence the decision to perform the gold standard? A common shortcoming of studies of diagnosis is to perform the gold standard test on only a portion of the enrolled patients. For example, one might study a diagnostic test such as a ventilation – perfusion scan that is used to diagnose pulmonary embolism. It will be impossible to calculate a false positive test rate for ventilation – perfusion scanning if the gold standard is not performed in patients with a positive ventilation – perfusion scan.

#### What are the results?

- 1) What likelihood ratios were associated with the range of possible test results? Likelihood ratios are becoming the preferred method for expressing the utility of diagnostic tests.

#### How can I apply the results to patient care?

- 1) Will the reproducibility of the test results and its interpretation be satisfactory in my clinical setting? The evidence is most useful if the diagnostic test under consideration is available to the patient with the clinical problem.
- 2) Are the results applicable to the patient in my practice? Does your patient resemble the subjects who were enrolled in the study?
- 3) Will the results change my management strategy? A diagnostic test is not relevant if it does not change patient management. For example, the anticoagulation protocol used for deep vein thrombosis is the same one used for pulmonary embolism. Therefore, there is usually no benefit gained by testing for pulmonary embolism if deep vein thrombosis has been diagnosed.
- 4) Will patients be better off because of the test? The ultimate utility of a diagnostic test is a determination of whether performing the test is associated with more patient benefit than harm. Harm may occur from performing the test (risk of allergic reaction from the contrast injected during a CT scan) or from any obligatory investigations arising from unexpected test results. For example, abdominal CT scans that are performed to assess the pancreas may pick up incidental tumors in the adrenal gland. These 'incidentalomas' are usually benign. However, they may have properties that are associated with malignant (cancerous) changes, which will ultimately oblige the clinician to proceed with further investigations that carry additional risk of harm such as a biopsy.

### **Critical Appraisal – Articles about Prognosis**

In assessing an article about a disease prognosis, three areas are considered.

#### Are the results valid?

- 1) Was the sample of patients representative? For the study to be valid, the patients in the sample should be typical of patients seen in clinical practice. A common problem is that sampling may take place in tertiary referral centres where researchers typically work. Cases referred for assessment at a tertiary referral centre are usually more severe or unusual than cases seen in the community. Studying these patients will give a biased assessment of prognosis.
- 2) Were the patients sufficiently homogenous with respect to prognostic risk? Studies of prognosis are most useful if the results of the trial apply to all members of the study sample. This is best accomplished if those enrolled in the study are at the same point in the course of their disease.
- 3) Was follow-up sufficiently complete? Losses to follow-up will reduce the ability of a study to determine patient outcome. For example, patients who do not return for a follow-up appointment owing to death will bias the results toward a more favorable prognosis than the true prognosis for the disease.

- 4) Were objective and unbiased outcome criteria used? To be valuable, the outcome under consideration needs to be standardized and reproducible. Mortality offers little challenge whereas measures of disability require more judgment and complex measurement.

#### What are the results?

- 1) How likely are the outcomes over time? This refers to the point estimate associated with the outcome of interest.
- 2) How precise are the estimates of likelihood? Once again, confidence intervals associated with the point estimate may be constructed. Narrow confidence intervals are desirable.

#### How can I apply the results to patient care?

- 1) Were the study patients and their management similar to those in my practice? This is an issue of generalizability of the study.
- 2) Was the follow-up sufficiently long? Patients may be interested in longer-term prognosis than what was assessed in the study.
- 3) Can I use the results in the management of the patients in my practice? Patients may vary with respect the outcome that is of most interest to their situation. For example, for an elderly patient mortality may be less important than the ability to live independently.

#### **Critical Appraisal – Articles about Harm**

Articles of assessing harm are likely to be observational in nature because it is usually not ethical, financially, or logistically feasible to control the degree to which a subject is exposed to a potential or known risk factor for disease. In assessing an article about harm, three areas are considered.

#### Are the results valid?

- 1) Did experimental and control groups begin with a similar prognosis?
- 2) Did the investigators demonstrate similarity in all known determinants of outcome; did they adjust for any differences in the analysis? Commonly, patient age and gender needs to be similar in any group being compared. Most disease outcomes will be influenced by these factors. If the groups are not similar, statistical techniques can be used to account for the effect of such variables.
- 3) Where exposed patients equally likely to be identified in both the disease and non-diseased groups? To reduce bias, it is important that the determination of exposure status be accurate.

- 4) Did experimental and control groups retain a similar prognosis after the study started?
- 5) Were outcomes measured in the same way in the groups being compared? To reduce bias, it is important that disease was measured in the same way in the groups.
- 6) Was follow-up sufficiently complete? Losses to follow-up risks biasing the outcome. For example, patients who do not return for a follow-up appointment owing to death will bias the results toward a more favorable prognosis than the true prognosis for the disease.

#### What are the results?

- 1) How strong is the association between exposure and disease? This refers to the point estimate associated with the outcome of interest.
- 2) How precise are the estimates of likelihood? Once again, confidence intervals associated with the point estimate may be constructed. Narrow confidence intervals are desirable.

#### How can I apply the results to patient care?

- 1) Were the study patients similar to the patients under consideration in my practice?
- 2) Was duration of follow-up adequate?
- 3) What was the magnitude of the risk?
- 4) Should I attempt to stop the exposure?

The above offers commonly used templates. Other critical appraisal templates are available including those to assess articles differential diagnoses, decision analyses, cost analyses, systematic reviews, and clinical practice guidelines.

## **UNIT 6: RESEARCH METHODS STUDY NOTES**

The purpose of the research methods unit is for the student to gain an understanding of the processes involved in clinical research. Students currently involved in research are provided time to further their research agendas. Clinical Research is emphasized because of the highly clinical orientation associated with undergraduate and post-graduate training. This unit builds on the knowledge and skills learned in units 1 through 4.

### **What is Clinical Research?**

A Panel of the U.S. National Institutes of Health offered a definition of clinical research in 1995. "Clinical research refers to investigations where the intact human is the unit of observation. Any material of human origin such as tissues, specimens, and cognitive phenomena may be observed in clinical research. However, the investigator (or colleague) directly interacts with the human subjects. Areas of clinical research may include: mechanisms of human disease, therapeutic interventions, clinical trials, development of new technologies, epidemiological and behavioral studies, outcomes research and health services research."

"Excluded from this definition are in vitro studies that utilize human tissues but do not deal directly with patients. In other words, clinical or patient-oriented research is research in which it is necessary to know the identity of the patients from whom the cells or tissues under study are derived. The Panel recognizes that there is no definition of clinical research upon which there is total agreement."

### **The Role of the Physician in Clinical Research**

It is common for physicians to participate in clinical research whether they are primarily clinicians or scientists, work in referral centres or small communities, or are generalists or specialists. It is important that physicians have an understanding of the research process.

Participation in research can vary considerably. Some physicians serve as principle investigators or co-investigators who plan and execute many aspects of their research. Other physicians play small but important roles such as: recruiting patients, caring for patients involved in research, measuring outcomes, or making observations. All physicians collect data that is used for research. For example, in Canada for every patient encounter that is paid for by a third party (private or public insurance), a physician is obliged to accurately record a diagnosis and/or procedure code associated with the encounter. Researchers use this information for a variety of purposes such as establishing health care needs and priorities. Canadian physicians are also obliged to report the presence of selected diseases as specified by public health laws. Hence, all physicians are responsible for the collection of data used by health care researchers.

Physicians may also be asked to serve as research subjects. For example, the Physician's Health Study assessed the effects of aspirin on the incidence of cardiovascular disease. Physicians may be asked to serve as control subjects, and physicians are commonly asked to participate in marketing research.

In summary, physicians have and will continue to play important roles in clinical research irrespective of the type of practice they pursue.

### **The Scientific Method**

It might be useful to review the Scientific Method. The Scientific Method refers to the principles that guide scientific research. Whereas philosophy in general is concerned with the why as well as the how of things, science occupies itself with the latter question only. 'The scientific method is the best way yet discovered for winnowing the truth from lies and delusion' according to one author.

The simple version of the Scientific Method looks something like this:

- 1) Observe some aspect of the universe.
- 2) Develop a theory that is consistent with what you have observed.
- 3) Use the theory to make predictions.
- 4) Test those predictions by experiments or further observations.
- 5) Modify the theory in the light of your results.
- 6) Go to step 3.

Whether or not the Scientific Method is responsible for advances in science is associated with some controversy. Those interested in the history of scientific theory can enter the names Thomas Kuhn and Karl Popper into an internet search engine.

Despite opinions to the contrary, most medical research is patterned on the Scientific Method and includes:

- 1) Developing an interest in an area of medicine or health care in general and making observations about current limitations.
- 2) Formulating a research problem that broadly defines a health related problem. For example, "how can we reduce back injuries in the workplace?"
- 3) Research the topic and identify gaps in the current state of knowledge.
- 4) Develop a research hypothesis that addresses the knowledge gap. For example, 'can morning stretching exercises reduce the frequency of back injuries in the workplace in factory workers?'
- 5) Test the hypothesis using an appropriate research design, collect the data, and formulate a conclusion.
- 6) Build on what was learned; go back to step 3.

### **Introduction to the Research Proposal**

In this course, you are asked to complete a research proposal. Research proposals should be completed prior to initiating any medical research. The purpose of a research proposal is to describe the population, intervention (if any), and outcome associated

with a research project. Funding agencies rely primarily on the research proposal to allocate financial resources.

The specific requirements of the proposal are outlined in the Assignments and Examinations Section of the Core Document. What follows is a guide to help you to complete this and possibly future research proposals.

Your research proposal should include and address the following headings, which will be discussed in more detail later in this section.

- 1) Title
- 2) Purpose
- 3) Background
- 4) Specific aims
- 5) Study design
- 6) Measurement
- 7) Data handling and analysis
- 8) Ethics
- 9) Budget
- 10) Significance and relevance
- 11) Strengths and weaknesses
- 12) Appendices

No one template will adequately cover the needs of every proposal. Do not be alarmed if a subheading does not seem to apply to your situation. For example, in a case-control study of a rare disease, it may not be appropriate to use a sample size based on a statistical calculation when a convenience sample (using the cases available at the time) is more feasible.

Please do not attempt to complete the research proposal in one evening. Leave sufficient time to gather the articles required for the background section. Sometimes, articles need to be ordered from other libraries. Independent study times are available during the latter half of the course to work on this assignment.

### **Choosing a Topic for the Research Proposal**

There are only two guidelines regarding the topic of your research proposal. First, it is strongly encouraged that proposal relate to clinical research. Clinical research is defined earlier in this section. Second, do not submit a research proposal that has been used elsewhere. Otherwise, you are free to choose a topic from any field of medicine or health care. You may build on knowledge gained from your Critically Appraised Topic or other academic experience.

Sometimes choosing a topic is the biggest hurdle in doing research. When choosing a topic, consider your interests. Think about the subjects you've covered so far and what lays ahead in your syllabus. Questions asked by your family or friends as well as conversations with colleagues may be good sources of ideas. Go to the library and

browse textbooks, the internet, or current journals. Consider the various news outlets for the latest advances. You may wish to consider building on what has been done previously. For example, for any given intervention, there may be outcomes that were not previously considered. Alternatively, there may be an intervention used for one condition that has not been studied in another.

Once you've decided on a topic, consider the following issues. Is the problem is significant enough to warrant investigation? Are you sure that the problem has not been answered? Can the problem be solved using scientific methods?

It is important for you to consider what you may wish to do with your research proposal. If you are thinking about carrying out the research proposal at a future date, you'll need to consider feasibility of the project in terms of funding, need for supervision, ethics, and time commitment.

### Title

Research proposal titles should be limited to one sentence that refers to the population, intervention, and outcome under consideration.

### Purpose

The Purpose is a sentence or two that broadly describes the objective(s) of the research. For example, 'the purpose of this research is to assess the effects of phototherapy on depression in adult males with bipolar disorder.'

### Background

The Background should provide a concise review of the research problem. The objectives of the background are to convey the importance of the proposed research, identify where there are gaps in the current knowledge base, and to provide support in favor of the intervention and methodology planned in your research project.

The background should cover the following areas where appropriate:

- 1) a brief description of the health problem under consideration,
- 2) an indication of how the health problem affects individuals and/or how frequently the health problem is encountered in the population,
- 3) a description of what is known from previous research and where gaps in the current knowledge base lay (this may include a discussion of the weaknesses associated with research done so far),
- 4) a rationale for the criteria used to define the study sample,
- 5) a rationale for the proposed study design,
- 6) a rationale for the intended intervention and consideration of subject safety,

- 7) a rationale for the outcomes under consideration and tools used to measure any outcome.

Be sure to reference all articles cited in the background section.

Keys to a good background section include: being complete with respect to the issues (1-7) suggested above, having a logical flow, being sure to include recent developments, citing original research / primary sources, indicating where critical appraisal of previous research has found significant problems, and building the case for further study.

A common problem for students is to devote most of the background to the description of the disease at the cost of defending the rationale or methodology of the study.

### Specific Aims

This part of the research proposal is variably referred to as the 'statement of objectives' or 'research questions' or 'hypotheses'. This section may contain statements that define the objectives of the research project. It is here where the research priorities are confirmed for the reader.

A primary research objective should be determined and explicitly stated. Defining a primary research objective is helpful for many reasons. It helps protocol reviewers to understand the investigator's priorities. The primary research objective also sets the stage for the sample size determination and is important for the statistical analyses that follow. Most investigators will have offered a prediction regarding the relationship between variables associated with the primary research objective.

The secondary objectives usually relate to other important albeit less frequent or significant outcomes that will be evaluated. Sometimes, secondary objectives are variations in the intervention or selected sub-group analyses. There may or may not be a statement of hypothesis regarding the relationship between variables included in the secondary objectives. Also, sample size determinations usually do not consider secondary objectives.

Tertiary objectives may be stated if appropriate or desired. Tertiary objectives may relate to the collection of data for generating new hypotheses or the collection of data regarding infrequent outcomes.

Keys to a successful statement of specific aims include: using clear and consistent terminology, indication the population, intervention, and outcomes under consideration.

### Study Design

This section is variably referred to as the 'methods section' or it may be divided among its components. The study design section should cover the following areas:

- 1) identify the intended research design (case-control, clinical trial),

- 2) describe from where research subjects will be recruited,
- 3) identify inclusion and exclusion criteria for entry of subjects into the research project,
- 4) what will be done to the subjects in the experimental arm, including a description of any activities done prior to allocation
- 5) what will be done to subjects in the control arm (describe the baseline standard of care if applicable),
- 6) describe when and what measurements will be done.

It is important that the study design addresses the objectives of the research. It may be helpful to use diagrams or tables to help describe the course and conduct of the research proposal. The research should be designed to minimize bias. It needs to be feasible and ethical.

### Measurement

The measurement section should identify all variables that are being measured, indicate how each variable will be measured, and how each measurement will be interpreted. It is paramount that the independent and dependent variables associated with the primary (and where possible secondary and tertiary) research objective be identified!

For example, in a descriptive study of disability after accidental burn (independent variable), there needs to be indication of how disability (the dependent variable) will be measured and what is to define the presence or absence of disability.

### Data Handling and Analysis

The data handling and analysis section should include:

- 1) a description of how data will be stored,
- 2) an indication of how the data will be safeguarded (this is becoming more important since passage of the Health Information Act in Alberta and similar legislation elsewhere),
- 3) any preliminary or interim safety analyses intended,
- 4) a description of all intended descriptive statistics (measures of central tendency and dispersion for selected variable), and
- 5) a description of all inferential statistics (statistical tests, confidence intervals or any modeling techniques anticipated).

### Ethics

The requirements of the Ethics Section will vary but generally includes an indication that the investigators intend to work in accordance with principles outlined by certain

authorities. An example of such an authority is the Canadian Institutes for Health Research that developed a Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. The statement is available on-line (see the reference materials section of the Core Document).

Important ethical principles include: respect for persons (autonomy), non-maleficence (do not harm), beneficence (do good), and justice (all persons must be treated fairly).

The Ethics section should describe in a few paragraphs:

- 1) that informed consent will be obtained from the study subjects,
- 2) any alternatives available to the subject (other treatments or what is likely to happen if the patient declines to participate in the research),
- 3) any potential risks or benefits that the subjects may experience,
- 4) any cost to the subject,
- 5) what the subject should do in case of emergency.
- 6) A consent form is NOT required for this course.

### Budget

The Budget section should describe the financial aspects of the research protocol. For the purposes of your research proposal, include:

- 1) a listing of all people who will need to contribute to the research project (consider recruitment, measurement, secretarial support, statistical support) and a sentence describing what each person will contribute, and
- 2) a list of material needed (include any lab tests, diagnostic imaging, computer equipment, software, clinic space, hospital beds, etc). You only need to budget for interventions that are beyond usual patient care. For example, in studying a new drug for myocardial infarctions, it is not usually necessary to budget for any health care that would normally have been provided to the patient.

There is no need to determine the dollar value of the human resources and materiel required for your research project.

### Significance and Relevance

The Significance and Relevance section briefly states how the results of this study will impact on the health of those affected with the condition under consideration. A few sentences should be sufficient; much of this information has been detailed in the background section.

### Strengths and Weaknesses

In a few sentences, summarize the strengths of your proposal. Perhaps, the research

question or intervention is novel. Or, your study design is superior to published research. Take this opportunity to 'sell' your proposal to a potential funding agency. With respect to weaknesses, indicate how your proposal may have been constrained by the availability of subjects, human resources, technology, ethical considerations, or other reasons.

### Appendices

Append any information that you feel is relevant to the proposal. Some examples may include: a copy of the product monograph associated with medication used in the experimental group, a picture of the device being assessed (new scalpel or brace), or a copy of the measurement tool (disability scale, quality of life assessment scale).