Epidemiology & Biostatistics for Curious Clinicians

Mark D. Harris, MD, MPH
Family, Sports, Preventive, Aerospace, Occupational and Environmental Medicine
Objectives

• Understand what epidemiology is and why it is important to clinical medicine
• Understand the fundamental concepts of clinical epidemiology

Abnormality
Diagnosis
Frequency
Risk
Prognosis

Treatment
Prevention
Cause
Cost
Part 1 - Introduction
What is Epidemiology?

• Greek - “epi” (upon) “demos” (people) “logia” (study)

• Epidemiology deals with the relationship of the various factors which determine the frequencies and distributions of an infectious process, a disease or injury, or a physiologic state in a human community

• Clinicians identify and treat disease and injury in individuals, while epidemiologists identify and prevent them in populations
Cholera Deaths in London – 1854 epidemic

Map created by Dr. John Snow
Elements of Clinical Epidemiology (1)

• In most clinical situations the diagnosis, prognosis and results of treatment are uncertain for individual patients and must be expressed as probabilities
  • Ex. 5 year survival rates, % clearing AOM w/o antibiotics

• Probability for an individual patient is best estimated by referring to past experience with groups of similar patients
Elements of Clinical Epidemiology (2)

- Because clinical observations are made on people who are free to do as they please and by clinicians and testing procedures with varying skills and biases, the observations may be affected by systematic errors that can cause misleading conclusions.
- All observations, including clinical ones, are also influenced by chance.
“And it was so typically brilliant of you to have invited an epidemiologist.”
Example of Sources of Errors in Clinical Medicine

- A 43 year old Chinese female presents to the ER with burning chest pain, shortness of breath and a 3/6 “machinery” systolic murmur x 36 hours. She has an important meeting to attend and “must be out of the ER in 30 minutes”. The lab’s troponin analyzer is down for repairs x 4 hours.

- Sources of potential errors/bias - language, culture, “premenopausal women don’t get heart disease”, “burning chest pain = GERD”, can’t hear murmur, patient’s schedule, can’t do lab...
To avoid being misled, clinicians should rely on observations that are based on sound scientific principles, including ways to reduce bias and estimate the role of chance.
What clinical questions does epidemiology answer?

• Abnormality - Is the patient sick or well?
• Diagnosis - How accurate are the tests (clinical, imaging, laboratory) used to diagnose disease?
• Frequency - How often does a disease occur?
• Risk - What factors are associated with increased risk of dz?
• Cause – What conditions lead to disease? What are the pathogenetic mechanisms of disease?
• Treatment – How does treatment change the course of dz?
• Prognosis - What are the consequences of having a disease?
• Prevention - Does an intervention on well people keep disease from arising? Does early detection and treatment improve the course of the disease?
• Cost - How much will care for an illness cost?
Negative Outcomes of Disease

- Death
- Disease - Signs and symptoms
- Discomfort - Pain, nausea, dyspnea, etc.
- Disability - Impaired ability to perform activities of daily living
- Dissatisfaction - Emotional reaction to disease
- Destitution - Financial cost of illness to patients, families and society
Negative Outcomes of Disease
Biological vs. Clinical Outcomes of Disease (e.g. CAD)

- **Biologic**
  - Coronary stenosis
  - Ejection fraction

- **Clinical**
  - Congestive heart failure
  - Quality of life
  - Death

These outcomes may or may not be related, and clinicians must remember that patients focus on the clinical ones.
Quantitating Measures and Outcomes

• Science is strongest when measurements are quantitative.
  • Better confirmation
  • More precise communication (Does “likely” = 75% or more? Does “rarely” mean 25% or less?)
    • Different clinicians assign different numbers to these terms
• Clinical outcomes can be expressed in numbers
• Individual patients will either have an outcome or not, but populations will have outcomes in probabilities
Tools for Quantitating

- **Ratios** - expressing the relationship of two terms
  - 10 boys and 5 girls in a classroom is 10:5 or 2:1
- **Proportion** - expressing the relationship of one part to the whole
  - The proportion of male students is 66%
- **Rate** - The expression of the probability of occurrence of a particular event in a defined population during a specified period of time
Examples of Quantitative Measurements - Rates

- **Mortality** - Total number of deaths from all causes in one year
  Number of persons in population at mid-year

- **Age specific mortality**
  Total number of deaths from all causes in an age group in one year (children < 10 yo)
  Number of persons in the population in that age group at mid-year

- **Cause specific mortality**
  Number of deaths from a specified disease per year (ex. Lung cancer)
  Number of persons in the population at mid-year

- **Case fatality** (ratio)
  Number of individuals dying during a specified period of time after disease onset/dx
  Number of individuals with a specified disease

- **Proportionate mortality**
  Number of deaths from a specified disease (ex. cardiovascular diseases)
  Total number of deaths in the US at mid-year
It would be best to quantify all of our measurements with rates based on the entire population. Often we can’t because it is hard to count hundreds of thousands or millions of people. So, we take a sample of the population and calculate rates based on those numbers.
Populations and Samples

• A **Population** is a large group of people in a defined setting (city, state, etc.) or with a defined characteristic (age, sex, etc.)

• A **Sample** is a subset of the population it is selected from. They are used for studies because it is usually not possible to study entire populations.
Example of Populations & Samples

Population (ex. a city, an occupational group)

Sample 1

Sample 2
Other Introductory Concepts

- Bias
- Confounding
- Chance
- Validity
Bias

- A process at any stage of inference tending to produce results that are systematically different than the true values.
- Selection bias occurs when comparisons are made between groups of patients that differ in determinants of outcome other than the one under study (Antiemetics in ship crew vs. passengers).
- Measurement bias occurs when the methods of measurement are dissimilar among groups of patients (PFTs done in an asthma cases, but only history in controls).
Confounding

- Occurs when two factors are associated (“travel together”) and the effect of one is confused with or distorted by the effect of the other.

Example

- A study shows coffee drinking is associated with Acute Myocardial Infarction (AMI). Is this real?
- Cigarette smoking is associated with AMI
- Coffee drinking is associated with cigarette smoking
- Cigarettes confound the association between coffee and AMI
One study in South America found that malaria was more prevalent in males than in females. Was that association due to some factor inherent in male biology or was it related to something else?

<table>
<thead>
<tr>
<th></th>
<th>→</th>
<th>Male gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Outdoor occupation</td>
<td>↓↓→</td>
<td>↓↓↓</td>
</tr>
<tr>
<td></td>
<td>→</td>
<td>Malaria</td>
</tr>
</tbody>
</table>
Chance

• Observations about a disease are normally made on a sample of patients.
• These observations vary by chance, even in the absence of bias.
• For example, taking the systolic blood pressure of 100 normal patients will result in the most often value of 120. However, the majority of values will not be 120 but will vary around that number.
Validity of a Study

- **Internal** - Are the results of a study correct for the sample of patients being studied?
- **External** - How much do the results of a study hold true in other settings?
Part 2 - Abnormality
Abnormality

- It is often difficult to distinguish normal from abnormal
  - Is a murmur benign or dangerous?
  - Is a slightly elevated lab value indicative of a dangerous disease or just a normal variant?
- For example, medicine teaches that patients with a fasting serum glucose of 135 and up have diabetes. However, some patients with levels of 150 do not have DM, and some with levels of 120 do. This is normal.
Types of Data

• Nominal - categories without inherent order (marital status, race, etc.) Can be dichotomous (yes/no, alive/dead)
• Ordinal - categories with inherent order
  • Does not have strictly defined intervals between categories
  • Murmur grades 1-6, edema grades 1-4
  • S. Agree, Agree, Neutral, Disagree, S. Disagree
• Interval
  • Continuous - Take on any value (age, chemistries)
Validity (Accuracy)

• How well the data measure what they were intended to measure.

• **Content** - How well a measurement method includes all dimensions of what one intends to measure and nothing more
  - A pain scale measuring aching/burning but not nausea/itching?

• **Construct** - How consistent is one method with another measuring the same thing?
  - Diagnosing ETOH use/abuse with CAGE data vs. AUDIT data.

• **Criterion** - How well a method predicts a directly observable phenomenon
  - How well does APFT data predict soldier physical performance and health?
Validity (Accuracy) and Reliability (Precision) of Tests
Reliability (Precision)

• How well do repeated measurements of a stable phenomenon (by different people and instruments in different times and places) get similar results.
  • Range - how well does it measure extremes?
  • Responsiveness - how well does the measure change as conditions change?
  • Interpretability - e.g. how do you interpret a survey cf a physical measurement?
Causes of Variation

- Measurement - Instrument and observer
- Biologic - Within individuals and among individuals
- Total - variation is cumulative
Expressions of Central Tendency

- **Mean** – the sum of all observations divided by the number of observations.
  - \( \bar{X} = \frac{1}{n} \sum X_i \) where \( n \) is the number of observations and \( x \) is each individual observation

- **Median** – the “middle” observation is a set. If there are \( N \) observations in a sample and these observations are ordered from smallest to largest, the median is
  - 1. The \(((N+1)/2)\)th largest observation in \( N \) is odd \((N = 9, m = 5)\)
  - 2. The average of the \((N/2)\)th and \((n/2 + 1)\)th largest observation if \( N \) is even. \((N = 10, m = 5.5)\)

- **Mode** – the most often seen value in a set of observations.
Expressions of Dispersion

- Range - the lowest to highest value in a distribution
- Standard deviation - the absolute value of the average difference of individual values from the mean
- Percentile, decile (10, 20, 30...), quartile (25, 50, 75), etc.
The Normal Distribution

-3  -2  -1  0  1  2  3

0.0215  0.1359  0.3413  0.3413  0.1359  0.0215
Criteria for Abnormality

• Abnormal = Unusual
• Abnormal = Associated with disease
• Abnormal = Treatable
Distributions

• Distributions are mathematical models to describe different phenomena
• It is useful to approximate reality with mathematical models because you can analyze models but it is often darn hard to analyze reality
Different Kinds of Distributions

Hazard Functions in Parametric Models

- Exponential (Risk of heart attack in middle aged adults)
- Weibull $p>1$ (Risk of death after leukemia diagnosis)
- Weibull $p<1$ (Risk of complications after surgery)
- Log-normal (Risk of clinical disease after infection)
<table>
<thead>
<tr>
<th>First Variable</th>
<th>Second variable</th>
<th>Example</th>
<th>Appropriate test(s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continuous</td>
<td>Continuous</td>
<td>Sodium level and blood pressure</td>
<td>Pearson correlation coefficient, linear regression</td>
</tr>
<tr>
<td>Continuous</td>
<td>Ordinal</td>
<td>Weight and happiness</td>
<td>ANOVA</td>
</tr>
<tr>
<td>Continuous</td>
<td>Dichotomous unpaired</td>
<td>Urine spec gravity and gender</td>
<td>Student’s T test</td>
</tr>
<tr>
<td>Continuous</td>
<td>Dichotomous paired</td>
<td>Weight before and after treatment</td>
<td>Paired T test</td>
</tr>
<tr>
<td>Continuous</td>
<td>Nominal</td>
<td>Age and presence/absence of MI</td>
<td>ANOVA</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Ordinal</td>
<td>Happiness and level of education</td>
<td>Spearman or Kendall correlation coefficients</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Dichotomous unpaired</td>
<td>Level of education and gender</td>
<td>Mann-Whitney U test</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Dichotomous paired</td>
<td>CHF classification before and after intervention</td>
<td>Wilcoxon matched pairs</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Nominal</td>
<td>Satisfaction and ethnicity</td>
<td>Kruskal-Willis test</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>Dichotomous unpaired</td>
<td>Success/failure in treated vs. untreated groups</td>
<td>Chi square, Fisher’s exact</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>Dichotomous paired</td>
<td>Success/failure before and after treatment</td>
<td>McNemar chi-square</td>
</tr>
<tr>
<td>Dichotomous</td>
<td>Nominal</td>
<td>Presence or absence of CVA and gender</td>
<td>Chi-square</td>
</tr>
<tr>
<td>Nominal</td>
<td>Nominal</td>
<td>Occupation and blood type</td>
<td>Chi-square</td>
</tr>
</tbody>
</table>
Part 3 - Diagnosis
Diagnosis - Making Data Simple

- To make a diagnosis, doctors must reduce complex information to simple scales
  - The many gradations of a heart murmur reduced to a 1-6 scale
  - Making a disease dichotomous (yes/no)
- Everything we do in medicine is simplified like this - physical exams, lab tests, etc.
Possible Outcomes of a Test

- **True Positive** - the test is positive and the patient actually has the disease
- **True Negative** - the test is negative and the patient actually does not have the disease
- **False Positive** - the test is positive but the patient does not have the disease
- **False Negative** - the test is negative but the patient actually does have the disease
The 2 x 2 Table

<table>
<thead>
<tr>
<th></th>
<th>Disease +</th>
<th>Disease -</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test +</td>
<td>True Positive</td>
<td>False Positive</td>
</tr>
<tr>
<td>Test -</td>
<td>False Negative</td>
<td>True Negative</td>
</tr>
</tbody>
</table>
The Gold Standard

• Every test, study or intervention must be compared to another one to show how useful it is.
• The study that new tests are compared to is called the gold standard
• Example - throat cultures are the gold standard for the diagnosis of strep throat. Other tests (antibody, etc.) are compared to it.
Problems with the Gold Standard

- They are never 100% sensitive and specific.
- They may be less accurate in reality than newer tests and therefore underestimate the effectiveness of the new test.
- Example - ultrasound is better at diagnosis pancreatic cancer than xray but is newer. So if xray is used as the gold standard, positives by US will be considered false positive.
Problems with Testing

- Information on negative medical tests is much less complete in the medical literature.
- There are often limited test results on the nondisease, since providers don’t order tests on the nondisease.
- So, the best information is on true positives.
- Lack of objective standards of disease.
Lead Time Bias

When evaluating the effectiveness of the early detection and treatment of a condition, the lead time must be subtracted from the overall survival time of screened patients to avoid lead time bias. Otherwise early detection merely increases the duration of the patients' awareness of their disease without reducing their mortality or morbidity. Numerous cancer screening procedures were thought to improve survival until lead time bias was addressed.

- Disease 'A' is 100% fatal and lasts 7 years from onset to death
  - Disease 'A' patients are asymptomatic for 4 years
  - Disease 'A' median survival 3 years after diagnosis

**Screening Test** can identify Disease 'A' early
- Detection occurs 2 years before clinical diagnosis
- Disease 'A' still lasts 7 years before death
  - Disease 'A' is now asymptomatic for 2 years
  - Disease 'A' median survival 5 years after diagnosis

**Screening Test** falsely appears to prolong survival
- Screened and unscreened patients each survive 7 years
- Screened patients know diagnosis 2 years earlier
Sensitivity

- How likely is the test to be positive if the patient has disease?
- Best to rule out disease
- Most useful if negative

\[ \text{Sensitivity} = \frac{TP}{TP + FN} \]

- Sensitivity is an inherent characteristic of the test – it does not depend on disease prevalence
Specificity

• How likely is the test to be negative if the patient does not have disease?
• Rule in disease
• Most useful if positive

\[ \text{Specificity} = \frac{TN}{TN + FP} \]

• Specificity is an inherent characteristic of the test – it does not depend on disease prevalence
Cut Off Point

• Balances sensitivity and specificity
• Example - Using the diabetes example earlier, do you set the cut-off point for diagnosing diabetes at a fasting blood glucose of 135 (high sensitivity but lower specificity) or 150 (high specificity but lower sensitivity)?
Receiver Operator Characteristic (ROC) Curve

- Plot of sensitivity vs 1-specificity
- A test with a good balance of sensitivity and specificity will be on the shoulder (upper left corner) of the curve
Effects of Choosing Different Cut-Off Scores on Diabetes Diagnosis

FIGURE 4–3. A–G, Screening for diabetes in a hypothetical population with a prevalence of 50%. Effects of choosing different cutoff levels for a positive test. (See text.)
Notes on Tests

• Patients used to describe the properties of a test should be the same as patients the test will be used on.
• Sensitivity and specificity are independent of prevalence.
Bias

- Occurs when the sensitivity and specificity of a test are not established independently of the means by which the true diagnosis is established.
- Example - When determining the sensitivity and specificity of reading x-rays, the true diagnosis is established by using clinical history (gold standard). However, radiologists use the clinical information to read the x-rays.
- So, the test agrees better with the gold standard and makes it seem more useful than it actually is.
Values for sensitivity and specificity are estimated from small samples which are themselves affected by chance.
Predictive Values

- **Positive Predictive Value (PPV)** - Probability of disease in a patient with a positive test.
- **Negative Predictive Value (NPV)** - Probability of not having a disease when the test is negative.

- More specific = better PPV
- More sensitive = better NPV
Increasing Prevalence

• Higher prevalence improves PPV
• Estimate prevalence from the medical literature, local databases and clinical judgment
• Populations from referral clinics have a higher prevalence of their associated diseases
• Certain demographic groups have higher prevalences of certain diseases
• People with specific clinical characteristics have higher prevalences of certain diseases
Likelihood Ratio (LR)

- Prob of that +/- test result in people with disease
- Prob of same test result in people without disease
- Probability = Odds/(1+odds)
- Pretest odds x LR = Posttest odds
- Main advantage - make it easier to go beyond the simple and clumsy classification of a test result as easier normal or abnormal.
  - “With this test result, your odds of having this disease is XX”
Multiple Tests - Parallel

- Tests done all at once
- Used when rapid assessment is necessary (i.e., emergency patients)
- Increases the sensitivity and therefore the NPV
- Also useful when a clinician needs high sensitivity but has only relatively specific tests
- Example: CBC, Chem 20, UA, LFT, CXR all done at once on trauma patient in E.D.
Multiple Tests - Serial

- First test ordered - If positive, second test ordered - If positive, third test ordered - etc.
- Used when some tests are expensive
- Used less laboratory resources, but takes more time.
- Maximizes specificity and PPV
- Example: First H&P for CAD, then GXT, then angiogram
Assumptions

• The information from test 3 is independent from the information from test 2 is independent from the information from test 1
• Restated, subsequent tests do not duplicate known information
THE SCIENTIFIC COMMUNITY IS DIVIDED.
SOME SAY THIS STUFF IS DANGEROUS, SOME SAY IT ISN'T.
Part 4 - Frequency
Qualitative Descriptions of Frequency

• Words like “rarely” or “frequently” are often used to describe the probability of an event.

• These words are very nonspecific and have different meanings depending upon who is using them.
Quantitative Descriptions of Frequency - Prevalence

- The fraction (proportion) of a group of people possessing a clinical condition or outcome at a given point in time
- Point - Cases present at the time of the survey
- Period - All cases present during a defined period.
- Example - How many people had hypertension in June 2013?
Prevalence

Number of cases of a disease present in the population at a specified time
Number of persons in the population at that specified time
Quantitative Descriptions of Frequency - Incidence

- The fraction (proportion) of a group of people initially free of a given clinical condition or outcome who develop it over a given period of time.
- Example: How many people develop (new onset) hypertension in 2001?
Incidence Rate

Number of new cases of a disease occurring in a population during a specified period of time

Number of persons at risk of developing the disease during that period of time
### Characteristics of Incidence and Prevalence

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Incidence</th>
<th>Prevalence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Numerator</strong></td>
<td>New cases occurring during a period of time among a group initially free of disease</td>
<td>All cases counted on a single survey or examination of a single group</td>
</tr>
<tr>
<td><strong>Denominator</strong></td>
<td>All susceptible people present at the beginning of the period</td>
<td>All people examined, including cases and non-cases</td>
</tr>
<tr>
<td><strong>Time</strong></td>
<td>Duration of the period</td>
<td>Single point</td>
</tr>
<tr>
<td><strong>How Measured</strong></td>
<td>Cohort study</td>
<td>Cross sectional study</td>
</tr>
</tbody>
</table>
What is a Case? Defining the Numerator

• To interpret rates, we must know how a case is defined, because the criteria used to define a case can strongly affect rates

• Include only “definite” cases (using the gold standard) or also “probable” cases (using often cheaper and easier tests)

• Example - How do you define a case of malaria? Relapsing fever? Organisms on culture? Both? Something else?
What is the Population at Risk? Defining the Denominator

- A rate is only useful to the extent that the physician can decide to which kinds of patients the rate applies.
- Populations may be all people in a given area (Los Angeles), all people in a given job (military), all people in a given medical care setting (hospital), or all people with a combination of factors such as those above.
Portals of Entry into Humans

- Respiratory Tract
- Mouth
- Conjunctiva
- Scratch, injury
- Arthropod
- Alimentary Tract
- Urinogenital Tract
- Anus
- Skin
- Capillary
Sampling

- It is usually impossible to study an entire population, so instead you study a sample.
- Random - every individual in the population has an equal probability of being selected.
- Probability - every individual has a known (but not necessarily equal) probability of being selected (i.e. more of a specified race).
- Convenience - individuals in a given practice accessible to the investigators.
Relationship among Incidence, Prevalence and Duration of Disease

Prevalence = Incidence \times \text{Average Duration}

(P = I \times D)
Discovering Frequencies of Events - Studies

- Descriptive studies - describe person/place/time
  - Case reports
  - Case series
- Analytic studies - generate hypotheses
  - Cross-sectional (prevalence)
  - Case control (retrospective)
  - Cohort (incidence, prospective)
- Experimental studies - vaccine or clinical trials
Cohort Study

• Identify a population free of the event of interest, then follow them through time via periodic examinations to determine occurrences of the event.

• Cumulative incidence - rate of new events in a group of people of fixed size, all of whom are observed over a period of time

• Incidence density - measure the number of new cases occurring in an ever-changing population where people are under study and susceptible for varying periods of time (person-time at risk)
Bias in Cross-Sectional Studies

- Uncertainty about time - what came first, disease or exposure
- Studying “old” cases - patients with both short and long duration of disease will be represented. Prevalence rates will be dominated by those patients who are able to survive their disease without losing its manifestations. They may be significantly different than the other cases and skew results
<table>
<thead>
<tr>
<th>Study Group</th>
<th>Cohort Studies</th>
<th>Case Control Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exposed persons (a+b)</td>
<td></td>
<td>Persons with the disease (cases a+c)</td>
</tr>
<tr>
<td>Non-exposed persons (c+d)</td>
<td></td>
<td>Persons without the disease (controls, b+d)</td>
</tr>
<tr>
<td>Outcome Measurement</td>
<td>Incidence in the exposed a/(a+b) and incidence in the non-exposed c/(c+d)</td>
<td>Proportion of cases exposed a/(a+c) and proportion of controls exposed b/(b+d)</td>
</tr>
<tr>
<td>Measures of Risk</td>
<td>Absolute risk</td>
<td>Odds ratio</td>
</tr>
<tr>
<td></td>
<td>Relative risk</td>
<td>Attributable risk</td>
</tr>
<tr>
<td></td>
<td>Odds ratio</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Attributable risk</td>
<td></td>
</tr>
<tr>
<td>Temporal relationship between exposure and disease</td>
<td>Easily established</td>
<td>Sometimes hard to establish</td>
</tr>
<tr>
<td>Multiple associations</td>
<td>Can study associations of an exposure with several diseases</td>
<td>Can study associations of a disease with several exposures or factors</td>
</tr>
<tr>
<td>Length of Study</td>
<td>Generally long</td>
<td>May be short</td>
</tr>
<tr>
<td>Cost of Study</td>
<td>Expensive</td>
<td>Less expensive than concurrent</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relatively inexpensive</td>
</tr>
<tr>
<td>Population size needed</td>
<td>Relatively large</td>
<td>Relatively large</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Relatively small</td>
</tr>
<tr>
<td>Potential Bias</td>
<td>Assessment of outcome</td>
<td>Susceptible to bias both in assessment of exposure and of outcome</td>
</tr>
<tr>
<td>Best when</td>
<td>Exposure is rare, disease is frequent among exposed</td>
<td>Exposure is rare, disease is frequent among exposed</td>
</tr>
<tr>
<td></td>
<td>Disease is rare, exposure is common among diseased</td>
<td></td>
</tr>
<tr>
<td>Problems</td>
<td>Selection of non-exposed comparison group may be difficult</td>
<td>Selection of non-exposed comparison group may be difficult</td>
</tr>
<tr>
<td></td>
<td>Loss to follow-up</td>
<td>Loss to follow-up</td>
</tr>
<tr>
<td></td>
<td>Changes over time in criteria and methods</td>
<td>Changes over time in criteria and methods</td>
</tr>
<tr>
<td></td>
<td>Selection of appropriate controls is often difficult</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Incomplete information on exposure</td>
<td></td>
</tr>
</tbody>
</table>
Uses of Incidence and Prevalence

- Predicting the future - predict the probability of an outcome by comparing the patient with similar patients (cohort studies)
- Assigning probability that a patient has a condition
- Making comparisons to aid in decision making
Clinical Decision Analysis

- Create a decision tree
- Assign probabilities to chance nodes
- Assign utilities to the outcomes
- Calculate the expected utilities
- Select the choice with the highest expected utility
- Sensitivity analysis
Would this be a better way to teach Epi and Biostats?

**Dilbert**

**LET'S HAVE A PRE-MEETING BEFORE YOUR MEETING WITH OUR VICE PRESIDENT.**

**DON'T MENTION ANY PROBLEMS BECAUSE HE MIGHT TRY TO FIX THEM.**

**DON'T SAY ANYTHING ABOUT BUDGETS OR DEADLINES BECAUSE HE MIGHT REDUCE THEM.**

**LEAVE OUT THE TECHNICAL STUFF BECAUSE IT WILL ONLY CONFUSE HIM.**

**THAT LEAVES ME NOTHING TO TALK ABOUT.**

**HELLO... AND IN SUMMARY, ARE THERE ANY QUESTIONS?**

**WOW! THAT'S THE FIRST PRESENTATION THAT HASN'T MADE ME FEEL NAUSEATED OR DIZZY! GREAT JOB!**

**WHY DOES SUCCESS MAKE ME HATE HUMANITY?**

**THEY DESERVE IT.**
Part 5 - Risk
Risk Factors

- Characteristics that are associated with an increased risk of becoming diseased.
- Examples - genetics (HLA B27), environmental (drugs, depression, etc).
- Exposure to a risk factor means that a person has, before becoming ill, come in contact with or has manifested the factor in question.
- Amount of exposure can be characterized by ever exposed, current dose, highest dose, total cumulative dose, etc.
Difficulties in Recognizing Risk

• Long latency between exposure and disease
• Frequent exposure to risk factors - so familiar that it doesn’t seem dangerous (smoking)
• Low incidence of disease - lung cancer even in heavy smokers is less than 2 in 1000
• Small risk - many factors only confer a small risk
• Common disease with many known risk factors - less incentive to look for new risk factors
• Multiple causes and effects - rarely close one-to-one relationship between factor and disease
Uses of Risk - 1

- **Prediction** - Predict the occurrence of disease
- **Cause** - Just because risk factors predict disease, they do not necessarily cause disease. Risk factors that do not cause disease are called “markers”
- **Example** - HPV is a risk factor for cervical cancer and causes it, while HSV is a marker for cervical cancer but does not cause it.
Uses of Risk - 2

- Diagnosis
  - Presence of a risk factor increases the probability (prevalence) of disease among patients
  - Absence of risk factors can also help rule out a disease

- Prevention - if a risk factor causes a disease, removing it can prevent the disease even if we don’t know how it causes the disease (i.e., John Snow turning off the Broad St. water pump in the London Cholera epidemic of the 1800s)
CELL RESPONSE
- Lysis of cell
- Inclusion body formation or cell transformation or cell dysfunction

HOST RESPONSE
- Death of organism
- Classical and severe disease
- Moderate severity mild illness

Below visual change
- Viral multiplication without visible change or incomplete viral maturation
- Exposure without attachment and/or cell entry

Infection without clinical illness (asymptomatic infection)
- Exposure without infection

Clinical Disease
- Subclinical Disease

Discernible effect
Studies of Risk

- **Experimental** - researcher determines who is exposed (i.e., vaccine efficacy trial)
- **Observational**
  - **Cohort** - begin with exposure and follow for outcome
  - **Case control** - begin with outcome track back to exposure
Cohort Studies

- A group of people is assembled, none of whom have experienced the outcome of interest, but all of whom could experience it.
- On entry to the study, subjects are classified according to those characteristics that might be related to the outcome (risk factors).
- Subjects should be observed over a meaningful period of time in the natural history of the disease in question.
- All members should be observed over the full period of follow-up to see who develops the outcome.
Other Names for Cohort Studies

• Longitudinal - emphasizing that patients are followed over time
• Prospective - implying the forward direction in which patients are pursued
• Incidence - noting the basic measure of new disease events over time
Historical (Retrospective) Cohort Study

• Cohort identified from past records and followed to the present (or future)
• Data quality is usually a concern, since data used was collected for some other purpose (than this study) and so may be incomplete or inaccurate
• Example - a study looking at medical records for long term effects of immunizations may lack that data in some participants
## Cohort

### Advantages/Disadvantages

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>The only way of establishing incidence directly</td>
<td>Inefficient because many more subjects must be enrolled than experience the outcome of interest</td>
</tr>
<tr>
<td>Follows the same logic as the clinical question – If they get exposed, will they get the disease?</td>
<td>Expensive because of the resources necessary to study many people over time</td>
</tr>
<tr>
<td>Can elicit exposure without the bias of already knowing the outcome</td>
<td>Assesses the relationship between disease and exposure to only relatively few factors (those recorded at study outset)</td>
</tr>
<tr>
<td>Can assess the relationship between exposure and many diseases</td>
<td>Results not available for a long time</td>
</tr>
</tbody>
</table>
Comparing Risks

- The basic expression of risk is incidence
- Measures of effect - measures of the association between exposure and disease
LIFELONG SMOKERS HAVE A ONE-IN-TWO CHANCE OF DYING FROM SMOKING-RELATED DISEASE.

IT'LL NEVER HAPPEN TO ME.

THE ODDS OF WINNING THE POWERBALL LOTTERY ARE 80 MILLION TO ONE.

THIS COULD BE MY LUCKY DAY!

By Steve Kelley, The San Diego Union Tribune, Tribune News Service
# Measures of Effect

<table>
<thead>
<tr>
<th>Expression</th>
<th>Question</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attributable risk (risk difference)</td>
<td>What is the incidence of disease attributable to exposure?</td>
<td>$AR = I_E - I_N$</td>
</tr>
<tr>
<td>Relative risk (risk ratio)</td>
<td>How many times more likely are exposed persons than unexposed persons to have the outcome of interest?</td>
<td>$RR = \frac{I_E}{I_N}$</td>
</tr>
<tr>
<td>Population attributable risk</td>
<td>What is the incidence of an outcome in the population, associated with a risk factor?</td>
<td>$ARP = AR \times P$</td>
</tr>
<tr>
<td>Population attributable fraction</td>
<td>What fraction of disease in a population is attributable to exposure to a risk factor?</td>
<td>$AFP = \frac{ARP}{IT}$</td>
</tr>
</tbody>
</table>

$I_E = \text{Incidence in the exposed}, \ I_N = \text{Incidence in the unexposed}, \ IT = \text{Total incidence}$
## Cohorts and their Purposes

<table>
<thead>
<tr>
<th>Characteristic in Common</th>
<th>To Assess</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Age</td>
<td>Life expectancy for people age 70 (regardless of when born)</td>
</tr>
<tr>
<td>Date of Birth</td>
<td>Calendar Time</td>
<td>Tuberculosis rates for people born in 1910</td>
</tr>
<tr>
<td>Exposure</td>
<td>Risk factor</td>
<td>Lung cancer in people who smoke</td>
</tr>
<tr>
<td>Disease</td>
<td>Prognosis</td>
<td>Survival rate for patients with breast cancer</td>
</tr>
<tr>
<td>Preventive Intervention</td>
<td>Prevention</td>
<td>Reduction in incidence of pneumonia after pneumococcal vaccine</td>
</tr>
<tr>
<td>Therapeutic Intervention</td>
<td>Treatment</td>
<td>Improvement in survival for patient’s with Hodgkin’s disease given combination chemotherapy</td>
</tr>
</tbody>
</table>
Part 6 - Prognosis
What will happen to me?

• Prognosis is the prediction of a future course of a disease following the onset
  • The goal is to predict each patient’s future as accurately as possible
  • Avoid being vague when it is unnecessary and avoid being certain when it is misleading

• Things to know
  • General course of the illness
  • Prognosis in a particular case (life/death)
  • Effect on quality of life for the patient
Prognosis Studies

• A group of patients with a medical condition are assembled and followed forward in time, and clinical outcomes are measured

• Clinical course – prognosis of a disease under treatment (ex. Diseases causing symptoms)

• Natural history – prognosis of a disease not under treatment (ex. Asymptomatic diseases)
Zero Time

- Point in time at which we begin to study a cohort of patients for a disease
  - Onset of symptoms
  - Time of diagnosis
  - Beginning of treatment
- If this is not clearly identified and standardized, observed prognosis would depend on the particular mix of zero times in our study
Outcome Concepts

• Describing Disease Outcomes
  • Include all 5 D’s – death, disease, discomfort, disability, dissatisfaction, and destitution
  • In general, use clinical measures of outcome
  • If an outcome cannot be related to something patients will recognize, don’t use it to guide patient care
• Health-related quality of life – clinical activities should have a positive impact on how patients function and live
Differences Between Prognostic and Risk Factors

- For Acute Myocardial Infarction
  - Risk – Older, male, smoking, hypertension
  - Prognosis – Older, female, smoking, hypotension, CHF, arrhythmia

- Different outcomes – Development of disease vs. death or development of complications

- Different rates
  - Risk factors predict low probability events
  - Prognostic factors predict much higher probability events
Using Rates to Describe Prognosis

- 5 year survival - % of patients surviving 5 years from some point in the course of their disease
- Case fatality - percentage of patients with a disease who die of it
- Disease-specific mortality - # of people per 10,000 or 100,000 who die of a specific disease
- Response - % of patients showing improvement after an intervention
- Remission - % of patients entering a phase in which disease is no longer detectable
- Recurrence - % of patients who return to a disease after a disease-free interval
Survival Analysis

- Examines the likelihood, on average, that patients with a given condition will experience a specific outcome at any point in time
Bias in Cohort Studies

• Susceptibility bias – occurs when groups of patients assembled for study differ in ways other than the factors under study
• Assembly bias – groups being compared are not equally susceptible to the outcome of interest
• Survival cohorts – patients are chosen for a cohort because they have the disease and have survived.
  • They describe the past history of prevalent cases, not what one would expect in a true cohort following disease prognosis from the beginning
• Type of assembly bias
Migration Bias

• A form of selection bias
• It can occur when patients in one group leave their original group, dropping out of the study altogether or moving to one of the other groups under study
• If the dropouts occur randomly and in small numbers, no bias results
• If the reasons for dropping out are related to prognosis and affect one group more than another, the results can be heavily biased
Measurement Bias

• Possible if patients in one group stand a better chance of having their outcome detected than those in another group

• Minimize by
  • Ensure that those making the observations don’t know which group the patient belongs to
  • Set and follow rules to decide whether an outcome event has occurred
  • Strive to discover events equally in all groups
CLASS A: INAPPA{NT INFECTION FREQUENT
Example: Tubercle bacillus

CLASS B: CLINICAL DISEASE FREQUENT; FEW DEATHS
Example: Measles virus

CLASS C: INFECTIONS USUALLY FATAL
Example: Rabies virus

Inapparent  Mild  Moderate  Severe (nonfatal)  Fatal
Dealing with Selection Bias
Today's Random Medical News

According to a report released today...

Jim Borgman
The Cincinnati Enquirer
Kung Features Syndicate
Randomization

- Assign patients to groups in a way that gives each patient an equal chance of falling into one or the other group
- Occurs in the design phase of the study
- The only way to equalize all extraneous factors, including those we know about and those we don’t
Restriction

- Limit the range of characteristics of persons in a study
- Occurs in the design phase
- Example – Only study males or females, blacks, Hispanics, Asians or whites, or some combination
- This may limit generalizability
Matching

- For each patient in one group, select one or more patients with the same characteristics (except the one under study) for the comparison group.
- It is usual only possible to match for a few characteristics (age and sex, for example).
- This makes data analysis much harder.
- Effects of matched variables on outcomes cannot be studied.
- Occurs during the design of a study.
Stratification

• Compare rates within subgroups (strata) with otherwise similar probability of outcome
• Occurs during the analysis phase of the study
• Example – Studies often stratify by age since that is often a major confounding factor
Adjustment

- Simple – Mathematically adjust crude rates for one or a few characteristics so that equal weight is given to strata of similar risk
- Multiple – Adjust for many characteristics with computer regression models
- Sensitivity analysis (best/worst case) – Describe how different the results would be under the most extreme conditions of selection bias
Other Factors

- Generalizability – How well can this data be applied to populations outside the study population?
Summary

- Abnormality
- Diagnosis
- Frequency
- Risk
- Prognosis

- Treatment
- Prevention
- Cause
- Cost
After 2 days of Epi and Biostats
References

• Fletcher, Wagner - Clinical Epidemiology: The Essentials
• Jekel, Katz, Elmore, Wild - Epidemiology, Biostatistics and Preventive Medicine