FUNDAMENTALS OF EPIDEMIOLOGY AND BIOSTATISTICS I

EPI 546

Spring 2016

Course Pack

(Course Protocol can be found on EPI 546 D2L website)

TABLE OF CONTENTS

See D2L for Syllabus Information
EPI 546 Course Overview, Requirements, & Evaluation
Faculty & Staff Contact Information

LECTURES

Last Updated 11/12/15

EPI 546: FUNDAMENTALS OF EPIDEMIOLOGY AND BIOSTATISTICS
Course Policies for 2016

Instructors:

East Lansing (EL)

Course Director:
Mathew J Reeves, BVSc, PhD (MR)
Professor
Department of Epidemiology and Biostatistics, CHM
East Lansing, MI 48824
reevesm@msu.edu

East Lansing Curriculum Assistant:
Dorota Mikucki
CHM-Office of Preclinical Curriculum
A- 112 X Clinical Center
East Lansing, MI 48824
Phone: (517) 884-1859
Dorota.Mikucki@hc.msu.edu

Grand Rapids (GR)

Assistant Course Director
Jeff Jones, MD (JJ)
CHM West Michigan
Jeffrey.jones@spectrum-health.org

Grand Rapids Curriculum Assistant:
Candace Obetts
CHM-Office of Preclinical Curriculum
15 Michigan Street, NE; #367
Grand Rapids, MI 49503
Office: (616) 234-2631
candace.obetts@hc.msu.edu

Information: For general course administrative questions contact Dorota Mikucki (if EL campus) or Candace Obetts (if GR campus). Students are encouraged to post content specific questions on the D2L Discussion Board. Either Dr. Reeves or Dr. Jones will post answers on a regular basis. Students are expected to check these postings on a regular basis. To arrange a face-to-face meeting with either Dr. Reeves or Dr. Jones, please contact by e-mail.

Required:

1) **Course pack,** referred to as **CP**
2) **Textbook:**
   Clinical Epidemiology: The Essentials, Fifth Edition, by Fletcher RH, and Fletcher SW. Lippincott, Williams & Wilkins 2014, Baltimore. ISBN 978-1451144475 (a.k.a. FF text in the notes).
   This book is available online at MSU library:
   http://libguides.lib.msu.edu/c.php?g=95640&p=624454
   Look under Year 1 EPI546.
   This book is also required for EPI 547 (Fall 2nd year).

3) **Optional Text: Users' Guides** to the Medical Literature: Manual of Evidence Based Clinical Practice (JAMA & Archives Journals), 2nd edition, by Gordon Guyatt, Drummond Rennie, Maureen Meade, Deborah Cook. McGraw-Hill Professional; May 21, 2008.

This book is available online at MSU library:
http://libguides.lib.msu.edu/c.php?g=95640&p=624454
Look under Year 2 EPI547. (This text is used in Year 2 for the EPI-547 course)

Lecture Location: All East Lansing Lectures and Help Sessions are in **A133 Life Sciences;** Exam Reviews are in **B105 Life Sciences.**
All Grand Rapids Lectures are in **120 Secchia Center (NOTE: GR will be in 130 Secchia on January 21 and February 25).**

Note that attendance at Lecture 1 is mandatory.

Lectures, FLIPPED Lectures, Application Sessions, and Assigned Readings:

Subject	FF Readings	Date	Time
1. Introduction to Epidemiology (MR)	Chapter 1	Tu 1/12	10:00-10:50 am
2. Descriptive Statistics	Chapter 2	Thur 1/14	online only
3. Frequency Measures (MR)	Chapters 4 and 5*	Tu 1/19	10:00-10:50 am
4. Effect Measures (MR)	Chapters 4 and 5*	Th 1/21	10:00-10:50 am
5. Statistics I (JJ)	Chapter 10	Tu 1/26	10:00-10:50 am
6. Statistics II (JJ)	Chapter 10	Th 1/28	11:00-11:50 am
Application Session I (JJ/MR)	See D2L	Mo 2/1	10:15 am-12:05 pm
Mid-Term Exam	All Above	Th 2/4	8:00-9:30 am
7. Clinical Testing (JJ)	Chapter 3	Tu 2/9	10:00-10:50 am
8. Prevention (MJR)	Chapter 9	Th 2/11	10:00-10:50 am
9. The Randomized Trial (JJ)	Chapter 8	Tu 2/16	9:00-9:50 am
10. XS, Cohort Studies (JJ)	Chapters 5 & 7**	Th 2/18	9:00-9:50 am
11. Case Control Studies (MR)	Chapter 6	Wed 2/24	8:00-8:50 am
Application Session II (MR/JJ)	See D2L	Th 2/25	8:00-9:50 am
12. Review/Help Session (MR/JJ)	All Above	Th 3/3	8:00-8:50 am
Final Exam	All Above	Fr 3/4	8:00-10:30 am

* Chapter 5 readings p 85-88
** Chapter 7 readings p 105-109 and 116-123

N.B – Some lectures will be given using a flipped format indicating that in-class time will be used for problems/discussion. Students should review required readings and the lecture slides beforehand and listen to the online lecture if they plan to participate in these sessions. These sessions will not be recorded on Mediasite.

Exam Schedule:
Mid-course exam worth 1/3 of the final marks will be on **Thursday, February 4 from 8:00 am - 9:30 am** in A133 Life Sciences (EL) and in 120 Secchia Center (GR). Final Examination will be on **Friday, March 4 from 8:00 am - 10:30 am** in A133 Life Sciences (EL) and in 120 Secchia Center (GR).

Overall Course Objective:
This course introduces the key concepts, definitions, vocabulary and applications associated with Clinical Epidemiology and Evidence-based Medicine that are fundamental to clinical practice and the critical appraisal of the medical literature.

Specific Objectives:

- Understand what clinical epidemiology is, and its relevance to the clinical practice of medicine through Evidence-based Medicine.

- Understand how clinical information is used to define “abnormal” vs. “normal.” Distinguish between validity and reliability. Understand the sources of variability (intra- vs. inter-observer/rater, biologic vs. measurement) and how they are quantified (standard deviation, correlation, kappa). Understand the rationale for sampling. Random vs. systematic error. Understand the different data classifications, data distributions, measures of central tendency and dispersion.
- Understand how to quantify uncertainty (probability and odds); ratios, proportions and rates; understand the definition, calculation, identification, interpretation, and application of measures of disease occurrence (prevalence, cumulative incidence, incidence-density, mortality, case-fatality); make probabilistic estimates about risk (absolute vs. relative); understand the relationships between incidence, duration and prevalence.
- Understand the definition, calculation, identification, interpretation, and application of measures of effect (RR, RRR, AR, ARR, ARI, NNT, NNH, PAR, PARF, OR); distinguish between relative and absolute differences; understand the relationship between baseline risk and ARR; describe the risks and benefits of an intervention through NNT and NNH measures; understand the distinction between the OR and RR, and recognize which study designs each measure can be applied.
- Understand the basic principles of medical statistics (hypothesis testing vs. estimation, confidence intervals (CI) and p-values, clinical significance and statistical significance); define and interpret p-values, point estimates, CIs, Type I and II errors; understand the determinants of power and sample size; on a conceptual level understand what multivariable analysis does (statistical control of confounding and interaction).
- Understand how to master the science and art of diagnosis and clinical testing; understand the definition, calculation, identification, interpretation, and application of measures of diagnostic test efficacy (sensitivity, specificity, predictive values); understand the importance of the gold standard, prevalence, and Bayes’ Theorem; understand the calculation and interpretation of ROC curves.
- Understand the different approaches undertaken to prevent disease (primary, secondary, tertiary), especially early detection through screening. Understand key concepts of population-level vs. individual-level prevention, mass screening vs. case-finding, Pre-clinical phase, and lead time. Understand difference between experimental vs. observation studies of screening efficacy and the importance of Lead-time bias, Length-time bias, and Compliance bias; criteria to assess feasibility of screening; understand the relative benefits and harms of screening.
- Understand the architecture of experimental (RCT) and observational study designs (Cross-sectional, cohort, case-control) along with their respective strengths and weaknesses; internal vs. external validity; understand the concepts of “bias” (selection, confounding, and measurement).
- Understand the design and key methodological steps of the RCT; understand the rationale for concealment, randomization and blinding and the biases each control; distinguish between loss-to-follow-up, non-compliance, and cross-overs; understand the different approaches used to analyze RCTs (ITT, PP, AT); distinguish between composite vs. individual measures, patient orientated vs. surrogate outcomes, pre-defined vs. post-hoc outcomes; best-cases vs. worst-case analyses.
- Understand the design and organization of a cohort study; distinguish between prospective vs. retrospective designs; recognize the common biases afflicting cohort studies - selection bias, loss-to-follow-up, generalizability; understand the difference between a “risk factor” and a “cause” of disease (association versus causation) and the uses of risk factor information.
- Understand the design and organization of a case-control study (CCS); distinguish between a case series, CCS, and a retrospective cohort; understand the principles used to select cases and controls; recognize the common biases afflicting CCS - selection bias, measurement bias (recall), confounding; list available mechanisms to control confounding; understand the role of matching; calculate and interpret the OR.

Course Material and Format:

All materials necessary for this course (i.e., glossary, lecture slides, course notes, and background readings (papers)) are in the course pack. All information in the course pack may be used to set examination questions. These materials along with practice questions and old examinations are also to be found on D2L. The course notes, glossary, and text book represent the primary sources of materials for the course. The lecture handouts (i.e., PowerPoint slides) are a summary of most, but not all, of the key concepts.

The course material will be demonstrated through a series of eleven 50-minute lectures (Lecture 2 will be presented only as an on-line lecture) and two 2-hour application sessions. The live lectures are intended to supplement the information in the course notes and the assigned readings in the textbook, but do not attempt to cover all of the material. Live lectures also serve as a venue for clarification and problem solving – students are encouraged to ask questions and to actively participate in in-class exercises. With the exception of the first lecture, attendance at the other lectures and applications sessions is voluntary. All live lectures are recorded and placed on Mediasite. In addition, for some of the subject areas, there are stand-alone pre-recorded lectures on D2L that cover all of the slides included in the lecture handouts.

For each of the 11 subject areas, several practice questions, generally in the same *format* as those on the mid-course and final exams (i.e., multiple choice, calculation based exercises, fill in the blank, and short answer format) are provided on D2L. These practice questions cover a broad range of difficulty – from easy to hard, and are **not** designed to be representative of the question *difficulty* that will be found on the two examinations (see further notes on practice examinations below). With respect to exam question wording and format, please note that we do not write the USMLE exam questions and so you should expect variability in terminology, vocabulary, and question format. This variability will occur on the board exams and elsewhere, thus it is an essential skill to be able to understand what a particular question is actually asking – even though the question may be worded differently than you would have liked it (or expected it) to be phrased. Having a firm grasp of the underlying concepts is therefore the best defense and it is this trait that the exam questions in this course are designed to test.

Application Sessions:
Two 2-hour "Application Sessions," each run by Dr. Jones and Dr. Reeves will provide a venue to apply the principles we have learned in the lectures to real world problems. These sessions will be interactive and students will be expected to work in small groups on short problems and present their answers to the class. Attendance is not required. The sessions will not be recorded. Reading materials necessary for these sessions will be found on D2L.

Expectations and Attendance Policy:
Apart from the first lecture, attendance is not required. However, attendance at the mid-course and final examinations is strongly advised.

Student Evaluation:
The course grading is based on 45 total marks (points) – to pass the course, students need to get 75% (i.e., >= 34 marks). Fifteen marks (33.3% of the total) will be based on a mid-course exam that will address material covered in all prior lectures. The final exam will be based on 30 questions. Of the 30 marks, at least 20 will be multiple choice with the remainder being calculation based, fill in the blank, and/or short answer format. The mid-course exam may also contain questions that are calculation based, fill in the blank, and/or short answer format. Two example mid-term exams and two example final exams are posted on D2L for self-assessment purposes. These are the ONLY officially endorsed practice examinations and the only exam questions on D2L that are designed to be representative of the *question difficulty* of the 2 examinations that will be used in this course. It is the policy of the EPI 546 course to keep the final exam secure. Students wishing to discuss the answers to specific questions should make an appointment with the course directors after receiving their final mark.

Those students not achieving the 75% benchmark will be given a CP grade and will be required to take the remediation exam.

Remediation Exam:

The remediation exam will be offered on **Tuesday, May 10, 2016 from 8:00 am - 10:30 am**, rooms to be announced, contact Dorota Mikucki (EL) or Candace Obetts (GR) for further details. The format of this exam will be similar to the final exam and the grading will be based on just the 30 questions included in the remediation exam (i.e., the scores from the mid-course exam will not be used).

Students who pass the remediation exam (i.e., >= 75%) will be given a CP/P grade. Those that fail the remediation exam will have their grade changed to CP/N and will have the opportunity to take the N remediation examination at the end of Block I. Students who fail the N remediation examination will be required to enter the Shared Discovery Curriculum.

Excused Absences and Make-Up Exams:

Students need to follow the Absence Policies of the College if they miss a required experience for any reason.

If illness, emergency, or other compelling circumstance makes it impossible for you to attend an examination session students should immediately fill a form entitled "**Request for Approval of Absence from an Examination or Required Experience,"** (found in D2L), and submitted by email to the appropriate address listed below:

East Lansing students submit to absencEL@msu.edu

Grand Rapids students submit to absencGR@msu.edu

If you are granted an excused absence from a course exam, your next step is to contact one of the administrative staff-persons below who will let you know the date, time, and place of the makeup exam.

Dorota Mikucki (East Lansing)	Dorota.Mikucki@hc.msu.edu	517-884-1859
Candace Obetts (Grand Rapids)	candace.obetts@hc.msu.edu	616-234-2631

Lecture 1 is mandatory; students who miss this lecture will be required to complete a make-up assignment that will be assigned by the two course assistants.

Directed Study Groups (DSG):
Supplemental instruction for this course is available free of charge. Students are encouraged to use these services which are led by epidemiology PhD graduate students. The sessions generally meet once a week for 2 hours or so. Please contact Veronica Miller (veronica.miller@hc.msu.edu), A139 LS (EL) or Renoulte Allen (Renoulte.allen@hc.msu.edu), Room 371 (GR) for more details.

Student feedback on course and instruction:
Forms on which to evaluate the course will be available at its conclusion and should be completed by 11:59 pm March 6th. Please take the time to provide feedback and your ideas on how to improve it.

Glossary of Practical Epidemiology Concepts - 2016

Adapted from the McMaster EBCP Workshop 2003, McMaster University, Hamilton, Ont.

Note that open access to the much of the materials used in the Epi-546 course can be found at http://learn.chm.msu.edu/epi/

Absolute Risk Reduction (ARR)
(or Risk Difference [RD] or Attributable Risk)

The difference in risks of an outcome between 2 experimental groups. Usually calculated as the difference between the unexposed or control event rate (CER) and the treated or experimental event rate (EER). Sometimes the risk difference is between 2 experimental groups.

$$ARR = CER – EER$$

Or

$$ARR = EER_1 – EER_2$$

The difference in risks between an exposed and unexposed group is also referred to as the *attributable risk* – that is, the additional risk of disease following exposure over and above that experienced by people not exposed (calculated as EER – CER).

Accuracy

Truthfulness of results or measurements. Requires comparison to known “truth”. Also referred to as validity.

Bias

Systematic error in study design which may skew the results leading to a deviation from the truth. A non-inclusive list of specific types can include:
Interviewerbias – error introduced by an interviewer’s conscious or subconscious gathering of selective data.

Lead-time bias – mistakenly attributing increased survival of patients to a screening intervention when longer survival is only a reflection of earlier detection in the preclinical phase of disease.

Recallbias – error due to differences in accuracy or completeness of recall to memory of past events or experiences. Particularly relevant to case control studies (CCS).

Referralbias – the proportion of more severe or unusual cases tends to be artificially higher at tertiary care centers.

Selectionbias – an error in patient assignment between groups that permits a confounding variable to arise from the study design rather than by chance alone.

Spectrum bias – occurs because of a difference in the spectrum and severity of disease between the population where the diagnostic test was developed and the clinical population that the test is applied to.. The disease subjects in the development population tend to be the “sickest of the sick” with few false negatives (and so Se is over-estimated) while the non-disease population tends to be the “wellest of the well” with few FP results (and so Sp is overestimated)

Verificationbias – when the decision to conduct the confirmatory or gold (reference) standard test is influenced by the result of the diagnostic test under study. Results in overly optimistic estimate of Se and an underestimate of Sp (a.k.a. work-up bias or test-referral bias).

Volunteer bias – people who choose to enroll in clinical research or participate in a survey may be systematically different (e.g. healthier, or more motivated) from your patients (a.k.a. responsebias).

Blinded or Masked

Blinded studies purposely deny access to information in order to keep that information from influencing some measurement, observation, or process (therefore blinding reduces information bias). “Double-blinded” refers to the fact that neither the study subject nor the study staff are aware of which group or intervention the subject has been assigned. Ideally everyone who is blinded or not should be explicitly identified (i.e, patients, doctors, data collectors, outcome adjudicators, statisticians).

Cochrane Collaboration

This international group, named for Archie Cochrane, is a unique initiative in the evaluation of healthcare interventions. The Collaboration works to prepare, disseminate, and continuously update systematic reviews of controlled trials for specific patient problems.

Co-intervention

Interventions other than the treatment under study. Particularly relevant to therapy RCTs' - readers should assess whether co-interventions were differentially applied to the treatment and control groups.

Concealment

A fine point associated with randomization that is very important. Ideally, you want to be reassured that the randomization schedule of patients was concealed from the clinicians who entered patients into the trial. Thus the clinician will be unaware of which treatment the *next* patient will receive and therefore cannot consciously – or subconsciously – distort the balance between the groups. If randomization wasn't concealed, patients with better prognoses may tend to be preferentially enrolled in the active treatment arm resulting in exaggeration of the apparent benefit of therapy (or even falsely concluding that treatment is efficacious). Note that concealment therefore reduces the possibility of selection bias (compare and contrast with blinding)

Confidence Interval (CI)

Clinical research provides a 'point estimate' of effect from a sample of patients; CIs express the degree of uncertainty or imprecision regarding this point estimate. CI represents a range of values consistent with the experimental data. In other words, CIs provide us with the 'neighborhood' within which the true (underlying and unknown) value is likely to reside. CIs and its associated point estimate help us make inferences about the underlying population.

The commonly used 95% CI can be defined as the range of values within which we can be 95% sure that the true underlying value lies. This implies that if you were to repeat a study 100 times, 95 of the 100 CIs generated from these "trials" would be expected to include the true underlying value. The 95% CI estimates the sampling variation by adding and subtracting 2 (or 1.96 to be exact) standard errors from the point estimate. The width of the CI is affected by inherent variability of the characteristic being measured, and the study sample size - thus the larger the sample size the narrower (i.e., more precise) is the CI. Finally, a useful rule to remember is that values outside of a 95% CI are statistically significantly different (at $P < 0.05$) from the point estimate. Acceptable formal definitions of the 95% CI in Epi-546 and 547 include:

i) The 95% CI includes a set of values which have a 95% probability or chance of including the underlying true value.

ii) The 95% CI is a measure of the precision surrounding the point estimate

Confounder or Confounding Variable

A factor that distorts the true relationship of the study variable of interest by virtue of being related to both the study variable and the outcome of interest. Confounders are often unequally distributed among the groups being compared. Randomized studies are less likely to have their results distorted by confounders because randomization should result in the equal balance of these factors at baseline.

Cost-Benefit Analysis (CBA)

Provides information on both the costs of the intervention and the benefits, expressed in monetary terms. Results are generally expressed as net benefits – benefits minus costs over a specific time period- or as the ratio of benefits to costs over a specific time period. Some of the CBA studies provide information only on the net savings, without providing details on the levels of costs and benefits. Other CBA-type studies provide information only on the monetary benefits or savings from an intervention without calculating the cost. A subset of these studies uses willingness-to-pay methods to calculate how much individuals and societies value the potential gains from an action/intervention.

Cost-Effectiveness Analysis (CEA)

Provides information on the cost of the intervention and its effectiveness, where effectiveness is not expressed in monetary terms but rather by a defined metric – generally the cost per life saved or case averted. CEA studies are in principle directly comparable if they use the same metric and the same methodologies in calculating costs.

Cost Only

Studies documenting costs of road traffic injuries without providing effectiveness or benefit information for actual or potential interventions.

Cost-Utility Analysis (CUA)

Similar to CEA, but the metric in the denominator is adjusted for quality of life or utility. CUA studies typically use Quality Adjusted Life Years (QALYs) or Disability Adjusted Life Years (DALYs) as their metric. QALYs and DALYs, both combine information on mortality and morbidity into a single indicator. Although many sourced make the distinction between CUA and CEA, in EPI-547 we generally refer to both types of analyses as CEA.

Cox regression model

A regression technique for survival analysis that allows adjustment for known differences in baseline characteristics between intervention and control groups when applied to survival data.

Diagnosis

The determination of the nature of a disease; a process of more or less accurate guessing.

Differential Diagnosis

A *probabilistic* listing of potential causes of a patient's clinical problem which can be ordered in a *probabilistic, prognostic,* or *pragmatic* fashion. Used to aid diagnostic decision-making.

Disability Adjusted Life Years (DALYs)

A negative measure of combined premature mortality and disability – i.e. the health gap between actual and potential health years of life. Death is defined as 1, and perfect health as 0. DALYs do not use interaction of types of morbidity, but rather add up disability weights from different conditions. DALYs are calculated using a population perspective; age weighting places a higher importance on individuals in prime productive age.

Discount rate

All types of economic evaluation of health conditions use some type of discounting to discount future benefits and costs, based on the principle that humans value benefits in the present more than they do benefits in the future. In theory, the discount rate should be equal to the real interest rate – i.e. the actual interest rate minus the rate of inflation. In practice, economic evaluation guidelines suggest using a rate of 3% annually.

Effectiveness

A measurement of benefit resulting from an intervention for a given health problem under conditionsofusualpractice. This form of evaluation considers both the efficacy of an intervention and its acceptance by those to whom it is offered. It helps answer "does the practice do more good than harm to people to whom it is offered?"

Efficacy

A measure of benefit resulting from an intervention for a given health problem under conditions of ideal practice. It helps answer “does the practice do more good than harm to people who fully comply with the recommendations?” (N.B. It is the job of RCTs’ to measure efficacy)

Event Rate (risk or CIR)

The risk or CIR of an event. Calculated as the proportion of a fixed population who develop the event of interest over a period of time. In a RCT design the terms controlled event rate (CER) and experimental event rate (EER) refer to the risks in the two comparison groups.

Evidence-Based Medicine

The conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients. The practice of evidence-based medicine requires integration of individual clinical expertise and patient preferences with the best available external clinical evidence from systematic research.

Generalizability (or external validity)

The extent to which the conclusions derived from a trial (or study) can be used beyond the setting of the trial and the particular people studied in the trial. Are the results applicable to the full population of all patients with this condition? Also referred to as External Validity. {See ‘internal validity’ and also ‘inference’ which deals with individualizing evidence to specific patients }.

Gold Standard
Reference Standard

An established method, or a widely accepted method, for determining a diagnosis. It provides a standard to which a new screening or diagnostic test can be compared. Most importantly in articles about diagnostic tests, the gold standard must be explicitly acknowledged and applied independently in a blinded fashion.

Hazard Ratio (HR)

The relative risk of an outcome (eg, death) over the entire study period; often reported in the context of survival analysis (Cox regression model). Has a similar interpretation to the relative risk.

Health

A state of optimal physical, mental, and social well being; not merely the absence of disease and infirmity (World Health Organization).

Health Outcome

All possible changes in health status that may occur in an individual or in a defined population or that may be associated with exposure to an intervention.

Heterogeneity

Differences between patients (clinical heterogeneity) or differences in the results of different studies (statistical heterogeneity).

Human Capital Approach

Defines costs and benefits in terms gains or losses in economic productivity. For an individual, injuries or deaths that are costed using the human capital approach include the theoretical future lost wages of the individual who died or was injured.

Inception Cohort

A designated group of persons assembled at a common time early in the development of a specific clinical disorder and who are followed thereafter. In assessing articles about prognosis it is critical that the inception cohort is well described in order to permit assessment of the homogeneity of the cohort.

Incidence rate

Number of new cases of disease occurring during a specified period of time; expressed either as a percentage (or proportion) of the number of people at risk (i.e., cumulative incidence rate [CIR]) or the number of new cases occurring per person time (i.e., incidence density rate - IDR).

Inference

To arrive at a conclusion. The act of taking information from published experience and individualizing to specific patients. The hierarchy and quality of available evidence significantly influence the strength of inference.

Intention-to-Treat Analysis

Analyzing patient outcomes based on which group they were randomized into regardless of whether they actually received the planned intervention. This analysis preserves the power of randomization, thus maintaining that important unknown factors that influence outcome are likely equally distributed in each comparison group. It is the most conservative but valid analytical approach for a RCT (compared to ‘as treated’ or ‘per protocol’ analyses). The term a *modified intention-to-treat analysis* is generally used to describe an analysis where the investigators excluded a small number of subjects from the pure ITT population (for example, patients who should not have been enrolled in the study or those who died shortly after enrollment of unrelated causes).

Internal validity

The degree to which inferences drawn from a specific study are accurate. Internal validity requires a careful assessment of the study’s methodology to determine whether the observed findings are accurate. Internal validity implies that apart from random error the study’s findings cannot be ascribed to a systematic error or bias; in other words the study does not suffer from confounding, selection or information bias to an important degree (judgment is required here). Thus RCT’s have high internal validity. Contrast with external validity or generalizability.

Kappa (K)

A measure of *reliability* or *agreement* between two raters for categorical or qualitative data (e.g., two physicians independently reading x-ray films). Kappa adjusts for the agreement that would be expected to occur due to chance alone, and is thus referred to a chance-corrected agreement. Kappa is preferred over other agreement measures such as the overall % agreement which is highly influenced by the prevalence of the condition being evaluated. Kappa ranges from -1 to +1. Values above 0.80 indicate excellent agreement, values 0.6-0.8 indicates substantial agreement, values 0.4-0.6 indicate moderate agreement, and values < 0.4 indicate fair or poor agreement.

Likelihood Ratio

A *ratio of likelihoods* (or probabilities) for a given test result. The first is the probability that a given test result occurs among people with disease. The second is the probability that the same test result occurs among people without disease. The ratio of these 2 probabilities (or likelihoods) is the LR. It measures the power of a test to change the pre-test into the post-test probability of a disease being present.

This ratio expresses the likelihood that a given test result would be expected to occur in patients with the target disorder compared to the likelihood of that same result in patients without that disorder. The LR for a given test result compares the likelihood of the result occurring in patients with disease to the likelihood of the result occurring in patients without disease. LRs contrast the proportions of patients with and without disease who have a given test result.

Matching

A deliberate process to make the study group and comparison group comparable with respect to factors (or confounders) that are extraneous to the purpose of the investigation but which might interfere with the interpretation of the studies’ findings. For example in case control studies, individual cases may be matched with specific controls on the basis of comparable age, gender, and/or other clinical features.

Median Survival

Length of time that one-half of the study population survives.

Meta-Analysis (MA)

A systematic review (SR) which uses quantitative tools to summarize the results.

Multivariable regression analysis

A type of regression model that attempts to explain or predict the dependent variable (or outcome variable or target variable) by simultaneously considering 2 or more independent variables (or predictor variables). Used to account for confounding and interaction effects. Examples include multivariable logistic regression (for binary outcomes) and multivariable linear regression (for continuous outcomes).

Non-inferiority trials

Trials undertaken with the specific purpose of proving that one treatment is no worse than another treatment (which is usually the current standard of care). Also includes equivalence trials.

Number Needed to Harm (NNH)

The number of patients who would need to be treated over a specific period of time before one adverse side-effect of the treatment will occur. It is the inverse of the absolute risk increase [ARI] (in the context of a RCT, the ARI is calculated as the EER – CER, where the event rates are adverse events, and by implication, adverse events are more common in the intervention, compared to control group).

$NNH = 1/ARI$ or $1/[CER*(RR-1)]$

– these equation are based on RCT or cohort study designs.

When faced with a CCS design that generates an OR the equation is:

$NNH = [CER(OR-1) +1] / [CER(OR-1)(1-CER)]$ [See page 227 UG]

Number Needed to Treat (NNT)

The number of patients who need to be treated over a specific period of time to prevent one bad outcome. When discussing NNTs it is important to specify the treatment, its duration, and the bad outcome being prevented. It is the inverse of the absolute risk reduction (ARR).

$NNT = 1/ARR$ or $1/[CER*(1-RR)]$

– these equation are based on RCT or cohort study designs.

When faced with a CCS design that generates an OR the equation is:

$NNT = [1 - CER(1-OR)] / [CER(1-CER)(1-OR)]$ [See page 226 UG]

Odds

A ratio of probability of occurrence to non-occurrence of an event

Odds = probability/1-probability

Odds Ratio (OR)

The *odds* ratio is most often used in case-control study (CCS) designs to describe the *magnitude or strength* of an association between an exposure and the outcome of interest (i.e. the odds of exposure in cases compared to the odds of exposure in the controls – it is therefore calculated as a ratio of odds). Because the actual underlying disease risks (or CIRs) in the exposed and unexposed groups cannot be calculated in a CCS design, the OR is used to approximate the RR. The OR more closely approximates the RR when the outcome of interest is infrequent or rare (i.e., <10%). As a measure of the strength of association between an exposure and outcome the OR has the same interpretation as the RR. The OR is calculated as the cross-product ratio (a.d/b.c) where a, b, c, and d represent the respective cell sizes.

Outcomes

All possible changes in health status that may occur in following subjects or that may stem from exposure to a causal factor or to a therapeutic intervention.

P-value

The probability of obtaining the value of the test statistic at least as large as the one observed, under the assumption that the H_0 is true (i.e. $P(data|H_0\ true)$). The smaller the p-value, the lower your degree of belief is in the null hypothesis being true. From a hypothesis testing standpoint the p-value is used to make a *decision* based on the available data. {note that increasingly the reporting of confidence intervals (CI) are preferred over p-values because they provide much more useful information – including the precision of the data and their possible clinical significance }.

Population attributable risk (PAR)

The incidence of disease in a population that is associated with a risk factor. Calculated from the Attributable risk (or RD) and the prevalence (P) of the risk factor in the population i.e.,

PAR = Attributable risk x P

OR, it can be calculated as:

PAR = Total Incidence minus Incidence in unexposed

Example: Thromboembolicdisease(TED)andoralcontraceptives(OC):
Incidence of TED in overall population = 7 per 10,000 person years
Incidence of TED in OC users = 16 per 10,000 person years
Incidence of TED in non-OC users = 4 per 10,000 person years
25% of women of reproductive age take OC
So, PAR = [16 – 4] x 0.25 = 3 per 10,000 person years. This represents the excess incidence of TED in the population due to OC use.

Population attributable risk fraction (PARF) (a.k.a etiologic fraction)

The fraction of disease in a population that is attributed to exposure to a risk factor. Under the assumption that the risk factor is a *cause* of the disease, it represents the maximum potential impact on disease incidence if the risk factor was removed. It can be calculated from the PAR as:

PARF = PAR/Total Disease Incidence

OR, it can be calculated directly from the RR and P as follows:

PARF = P(RR-1)/ [1 + P(RR-1)]

Example: Thromboembolicdisease(TED)andoralcontraceptives(OC):

PARF = PAR/Total Disease Incidence
PARF = 3 per 10,000/7 per 10,0000 = 43%

PARF = P(RR-1)/ [1 + P(RR-1)]
PARF = 0.25(4-1)/ [1 + 0.25(4-1)] = 43%. This represents the fraction of all TED in the population that is due to OC use.

Power

Ability to detect a difference between two experimental groups if one in fact exists (1 – beta).

Precision

A measure of variability in the point estimate as quantified by the confidence interval. Influenced by random error.

Predictive Value

PositivePredictiveValue (PVP) – proportion of people with a positive test who have disease.

NegativePredictiveValue (PVN) – proportion of people with a negative test who are free of disease.

Prevalence (P, Prev)

Proportion of persons affected with a particular disease at a specified time. Prevalence rates obtained from high quality studies can inform clinician's efforts to set anchoring pretest probabilities for their patients. In diagnostic studies, also referred to as prior probability.

Prognosis

The possible outcomes of a disease and the frequency with which they can be expected to occur.

Quality Adjusted Life Years (QALYs)

Combine morbidity and mortality into a positive measure of life lived, with death defined as 0 and perfect health as 1. QALYs are defined from the individual's perspective, and include interaction of different types of morbidity. QALYs using discounting of future years, generally at a 3% discount rate.

Randomization

Allocation of individuals to groups by a formal chance process such that each patient has an independent, equal chance of selection for the intervention group.

Relative Risk (Risk Ratio)

The relative probability or risk of an event or outcome in one group compared to another. In a RCT, it is calculated as the ratio of risk in the treatment or experimental group (EER) to the risk in the control group (CER), and is a measure of the *efficacy (or magnitude*) of the treatment effect. In a cohort study, where it is similarly used to express the magnitude or strength of an association, it is calculated as the ratio of risk of disease or death among an exposed population to the risk among the unexposed population.

$RR = EER / CER$

Relative Risk Reduction (RRR)
Relative Risk Increase (RRI)

The percent reduction in an outcome event in the experimental group as compared to the control group. It is the complement of RR or the proportion of risk that's removed by the intervention. Unlike the ARR the RRR is assumed to be a constant entity – that is, it is assumed not to change from one population (study) to another. It therefore represents a fixed measure of the efficacy of an intervention (contrast with the ARR which is responsive to changes in the baseline event rates).

$RRR = 1 – \text{Relative Risk}$
$RRR = CER – EER / CER \text{ x } 100$
{Or obviously … RRR = ARR / CER}

The relative risk increase (RRI) is the percent increase in an outcome in the experimental group as compared to the control group, calculated as

$RRI = RR – 1.$

The distinction between RRR and RRI can get confusion hence many people prefer to use the *Risk Difference* which is simply the absolute difference between the EER and CER.

Reliability

Refers to consistency or reproducibility of data; a.k.a. repeatability. Referred to as agreement when examining categorical data (see also Kappa). *Intra-rater* reliability refers to the consistency within the same observer or instrument. *Inter-rater* reliability refers to the consistency between two observers or instruments. It is important to distinguish reliability from validity.

ROC Curves

Receiver operator characteristic (ROC) curves plot test sensitivity (on the y axis) against 1- specificity (on the x axis) for various cut-points of a continuously distributed diagnostic variable. The curves describe the tradeoff between Se and Sp as the cut point is changed. Test that discriminate well crowd towards the upper north-west corner of the graph. ROC curves can be used to compare the discriminating ability of two or more tests by comparing the area under the curve (AUROC).

Sensitivity (Se)

The proportion of people with disease who have a positive test or P(T+|D+)

SnNout

When a test with a high sensitivity is negative, it effectively rules out the diagnosis of disease.

Sensitivity Analysis

A test of the stability of conclusions by evaluating the outcome over a range of plausible estimates, value judgments, or assumptions.

Specificity (Sp)

The proportion of people without disease who have a negative test or P(T-|D-).

SpPin

When a test is highly specific, a positive result can rule in the diagnosis.

Standards

Authoritative statements of minimal levels of acceptable performance or results, excellent levels of performance or results, or the range of acceptable performance or results.

Study Designs

1. CaseSeries – A collection or a report of the series of patients with an outcome of interest. No control group is involved.

2. CaseControlStudy(CCS) – Identifies patients who have a condition or outcome of interest (cases) and patients who do not have the condition or outcome (controls). The frequency that subjects are exposed to a risk factor of interest is then compared between the cases and controls. Because of the design of the CCS, disease rates cannot be directly measured (contrast this with the cohort study design). Thus the comparison between cases and controls is actually done by calculating the odds of exposure in cases and controls. The ratio of these 2 odds results in the odds ratio (OR) which is usually a good approximation of the relative risk (RR)

   Advantages: it is relatively quick and inexpensive requiring fewer subjects than other study designs. It is often times the only feasible method for investigating very rare disorders or when a long lag time exists between an exposure of interest and development of the outcome/disease of interest. It is also particularly helpful in studies of outbreak investigations where a quick answer followed by a quick response is required. Disadvantages: recall bias, unknown confounding variables, and difficulty selecting appropriate control groups.

3. CrossoverDesign – A method of comparing 2 or more treatments or interventions in which all subjects are switched to the alternate treatment after completion of the first treatment. Typically allocation to the first treatment is by a random process. Since all subjects serve as their own controls, error variance is reduced.

4. Cross-SectionalSurvey – The observation of a defined population at a single point in time or during a specific time interval. Exposure and outcome are determined simultaneously. Also referred to as a prevalence survey because this is the only epidemiological frequency measure that can be measured (in other words incidence rates cannot be generated from this design)

5. CohortStudy – Involves identification of two groups (cohorts) of patients who are defined according to whether they were exposure to a factor of interest e.g., smokers and non-smokers . The cohorts are then followed over time and the incidence rates for the outcome of interest in each group are measured. The ratio of these incidence rates results in the relative risk (RR) which quantifies the magnitude of association between the factor and outcome (disease). Note that when the follow-up occurs in a forward direction the study is referred to as a prospective cohort. When follow-up is done based on historical information it is referred to as a retrospective cohort).

   Advantages: can establish clear temporal relationships between exposure and disease onset. Are able to generate incidence rates . Disadvantages: control/unexposed groups may be difficult to identify, exposure to a variable may be linked to a hidden confounding variable, blinding is often not possible, randomization is not present. For relatively rare diseases of interest, cohort studies require huge sample sizes and long f/u (hence they are slow and expensive).

6. N-of-1Trial – When an individual patient undergoes pairs of treatment periods organized so that one period involves use of the experimental treatment and the other involves use of a placebo or alternate therapy. Ideally the patient and physician are both blinded, and outcomes are measured. Treatment periods are replicated until patient and clinician are convinced that the treatments are definitely different or definitely not different.

7. RandomizedControlledTrial – A group of patients is randomized into an experimental group and into a controlled group. These groups are then followed up and various outcomes of interest are documented. RCT's are the ultimate standard by which new therapeutic maneuvers are judged. Randomization should result in the equal distribution of both known and unknown confounding variables into each group. An unbiased RCT also requires concealment and where feasible blinding.

   Disadvantages: often impractical, limited generalizability, volunteer bias, significant expense, and sometimes ethical difficulties.

8. SystematicReview – A formal review of a focused clinical question based on a comprehensive search strategy and structured critical appraisal designed to reduce the likelihood of bias. No quantitative summary is generated however.

9. Meta-Analysis – A systematic review which uses quantitative methods to combine the results of several studies into a pooled summary estimate.

Survival analysis

A statistical procedure used to compare the proportion of patients in each group who experience an outcome or endpoint at various time intervals over the duration of the study (eg, death). (See also Cox regression analysis)

Survival curve

A curve that starts at 100% of the study population and shows the percentage of the population still surviving (or free of disease or some other outcome) at successive times for as long as information is available. (also referred to as Kaplan Meier survival curves)

Substitute or Surrogate Endpoints

Refer to study outcomes that are not immediately significant in clinical patient care. Substitute endpoints may include rates of biochemical changes (e.g., cholesterol, HbA_1C) while clinically significant endpoints are more clearly tied to events that patients and their doctors care about most (e.g., stroke, renal failure, death).

Validity

Truthfulness or believability of study conclusions or the extent to which a test actually measures what it is supposed to measure or accomplishes what it is supposed to accomplish. Simply put, "Does the data really mean what we think it does?" or "Can we believe the results?" A.k.a. accuracy. Validity implies the presence of a gold standard to which data can be compared to. See also internal validity and external validity

Willingness To Pay (WTP)

Costs benefit analysis, measures benefits as the aggregate sum of money that potential beneficiaries of an intervention would be willing to pay for the improvements that they would expect from that intervention. The WTP approach provides a methodology to place a financial value on potential gains from an action or intervention.

Need online access to these course materials?

Go to: http://learn.chm.msu.edu/epi/

Abbreviations

AR	Attributable risk
ARI	Absolute risk increase
ARR	Absolute risk reduction
CBA	Cost-benefit analysis
CEA	Cost-effectiveness analysis
CER	Control event rate
CI	Confidence interval
CCS	Case-control study
CIR	Cumulative incidence rate
CUA	Cost-utility analysis
EER	Experimental event rate
IDR	Incidence density rate
MA	Meta-analysis
NNH	Number needed to harm
NNT	Number needed to treat
OR	Odds ratio
P (or Prev)	Prevalence
PAR	Population attributable risk
PARF	Population attributable risk fraction
PVP	Predictive value positive
PVN	Predictive value negative
RCT	Randomized controlled trial
RD	Risk difference
ROC	Receiver operator characteristic curve
RR	Risk ratio or relative risk
RRR	Relative risk reduction
Se	Sensitivity
Sp	Specificity
SR	Systematic review
TER	Treatment event rate

EPI-546 Block I

Lecture 1: Introduction to Epidemiology and Biostatistics

What is epidemiology and evidence-based medicine and why do I need to know about it?

Mathew J. Reeves BVSc, PhD, FAHA Associate Professor, Epidemiology, CHM – EL

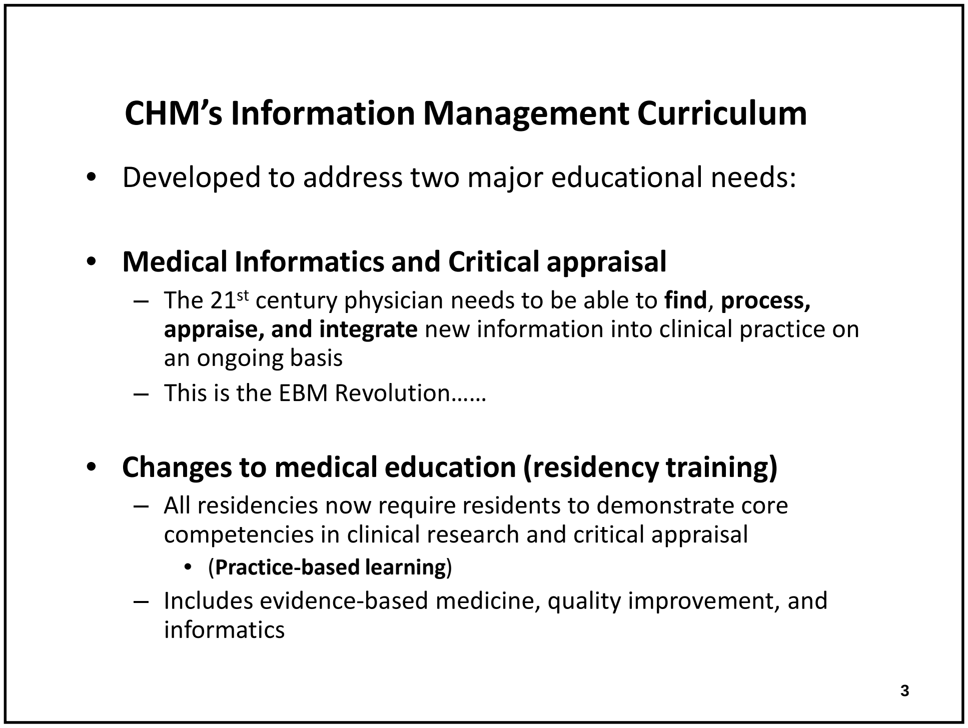
Jeffrey Jones, MD Emergency Medicine, CHM – GR

1

Medicine and the information age

- Medicine has become an **information intense** discipline
- The volume of new medical information is staggering
  - Example: MEDLINE (NLM, NIH)
  - Contains ~5,600 biomedical journals, 39 languages
  - Contains 19 million articles
  - 700,000 new articles added each year (2-4,000 a day!)
- Access to medical information has increased dramatically
  - Everyone is exposed to the media
  - Almost everyone has access to the internet
- Interest in medical information has increased exponentially
  - The media focus on “today’s medical research breakthrough”
  - Increasing awareness and demands of patients and payers'
  - Increasing demand for physician accountability

2

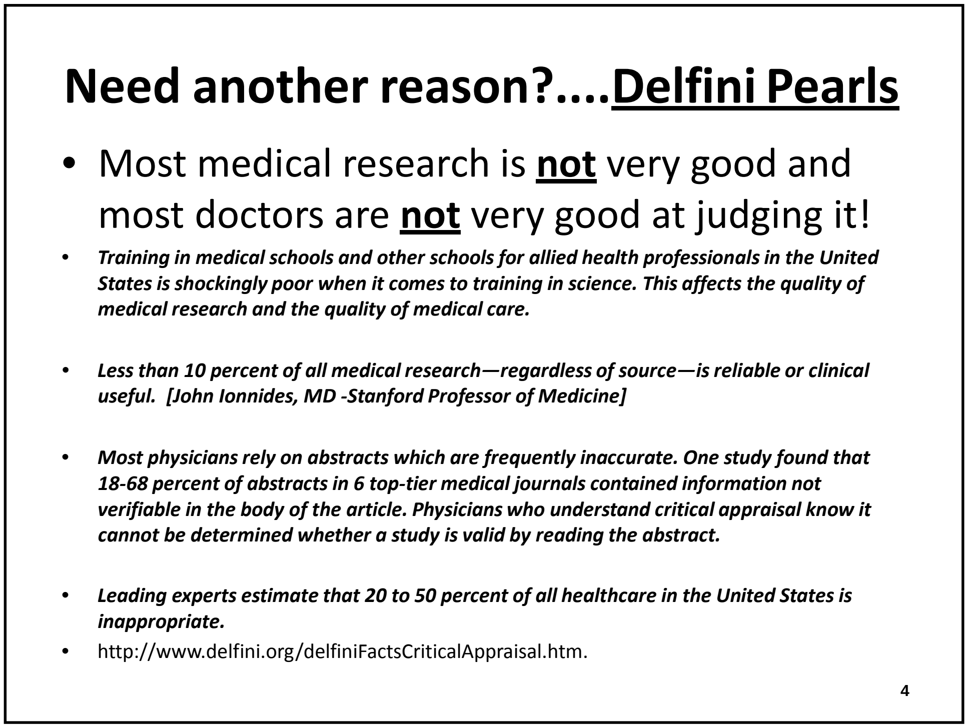


CHM’s Information Management Curriculum

- Developed to address two major educational needs:

- **Medical Informatics and Critical appraisal**
  - The 21st century physician needs to be able to **find**, **process, appraise, and integrate** new information into clinical practice on an ongoing basis
  - This is the EBM Revolution……

- **Changes to medical education (residency training)**
  - All residencies now require residents to demonstrate core competencies in clinical research and critical appraisal
    - (**Practice-based learning**)
  - Includes evidence-based medicine, quality improvement, and informatics



Need another reason?....<u>Delfini Pearls</u>

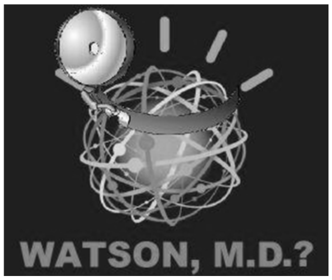
- Most medical research is **<u>not</u>** very good and most doctors are **<u>not</u>** very good at judging it!
- ***Training in medical schools and other schools for allied health professionals in the United States is shockingly poor when it comes to training in science. This affects the quality of medical research and the quality of medical care.***

- ***Less than 10 percent of all medical research—regardless of source—is reliable or clinical useful. [John Ionnides, MD -Stanford Professor of Medicine]***

- ***Most physicians rely on abstracts which are frequently inaccurate. One study found that 18-68 percent of abstracts in 6 top-tier medical journals contained information not verifiable in the body of the article. Physicians who understand critical appraisal know it cannot be determined whether a study is valid by reading the abstract.***

- ***Leading experts estimate that 20 to 50 percent of all healthcare in the United States is inappropriate.***
- http://www.delfini.org/delfiniFactsCriticalAppraisal.htm.

Need another reason?.... IBM’s Watson, MD – the end of traditional medicine? and traditional medical education?

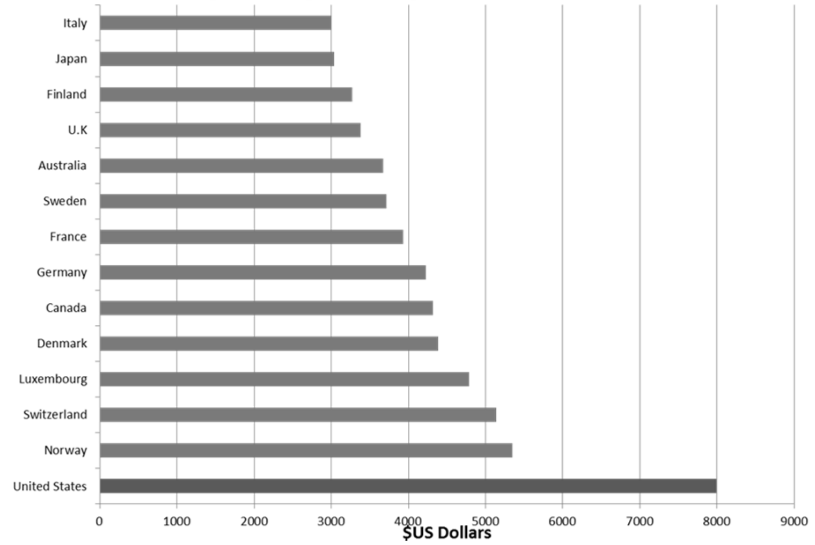


… Knowing medical facts will no longer be sufficient!

Need another reason?.... The U.S. Health System is “challenged” on many fronts…..

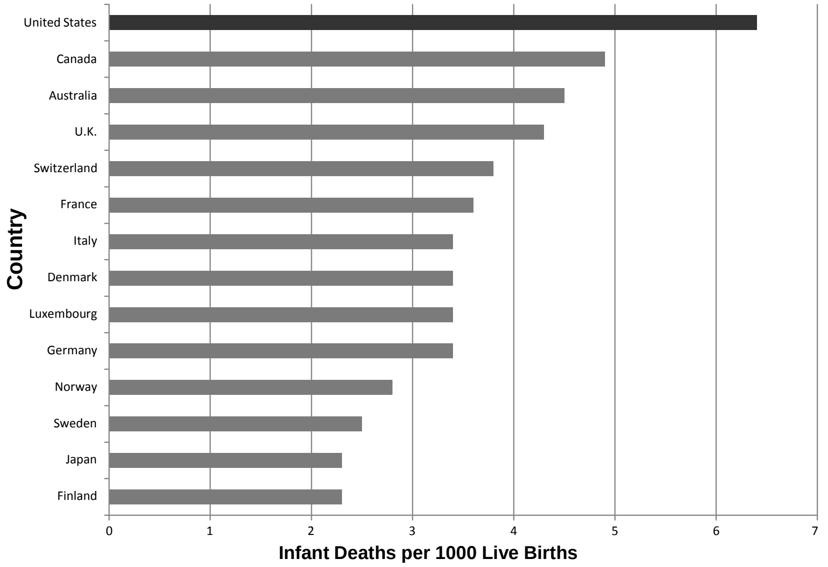
- Compared to other industrialized countries the U.S. health care system …..
  - – Is expensive: $2.6 trillion in 2010, 18% GDP
  - – Medical inflation (~5%/yr) is higher than growth of overall economy
  - – Does not produce the best outcomes
  - – Is not rated highly by its citizens or doctors
  - – Does not cover all of its citizens…. ~15% uninsured (~50 Million)
    - Source www.kaiseredu.org

Health Spending Per Capita in US Dollars, 2009.

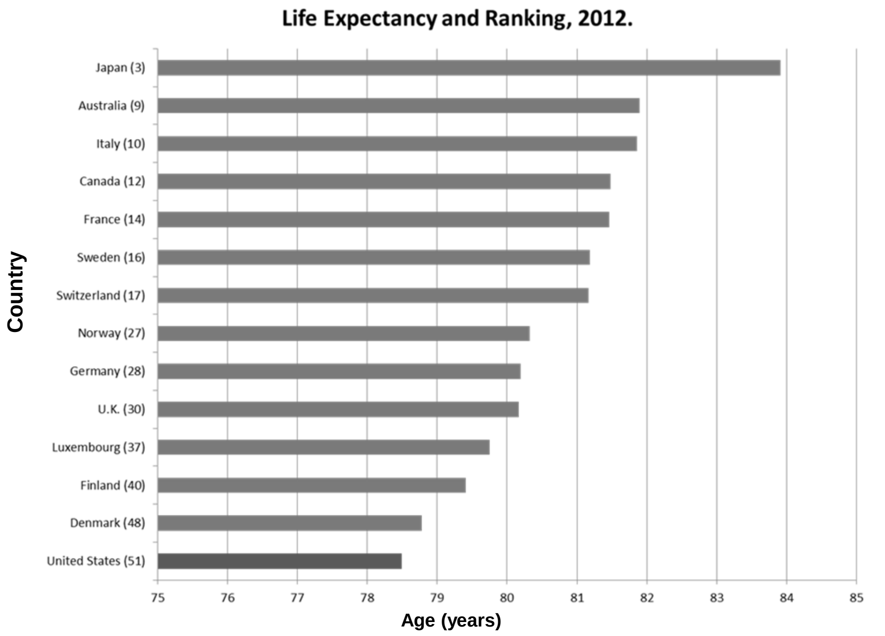


Source: http://www.oecd-ilibrary.org/social-issues-migration-health/total-expenditure-on-health-per-capita_20758480-table2

Infant Mortality Rate, 2010.

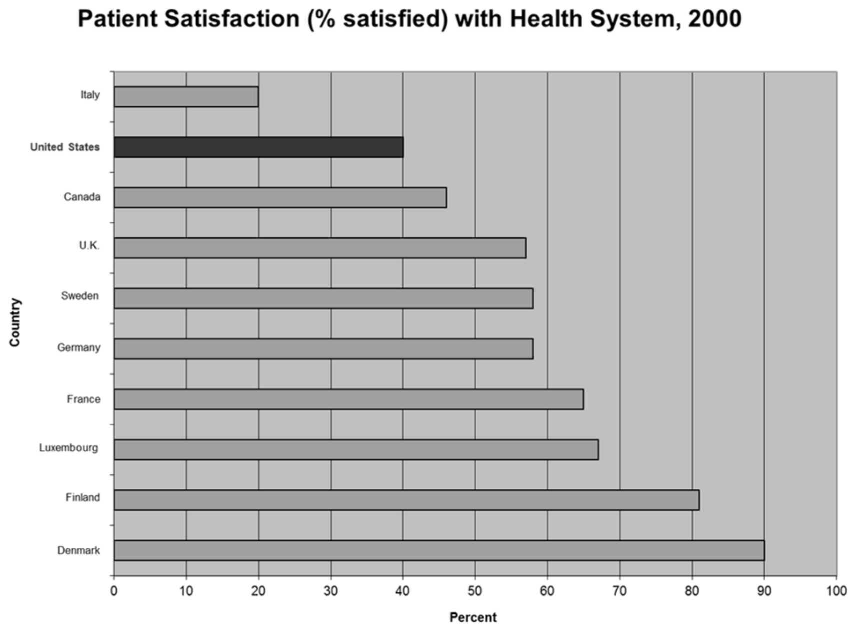


Sources: http://www.cdc.gov/nchs/deaths.htm
http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/database http://www.e-stat.go
http://www.abs.gov.au
/www.cia.gov



Source: https://www.cia.gov/library/publications/the-world-factbook/geos/mn.html

9

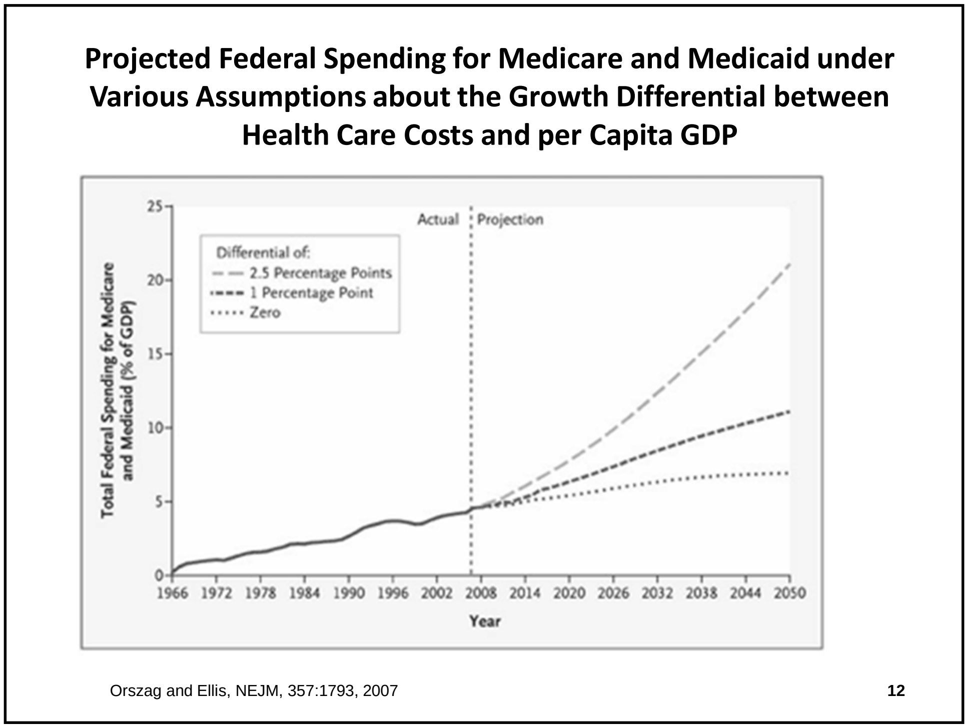


10

U.S. Health Care: A high cost but low quality industry.

- Congressional Budget Office, 2007:
  - "*The long-term fiscal condition of the US has largely been misdiagnosed. Despite the attention paid to the demographic challenges, such as the coming retirement of the baby-boom generation,* ***our countries financial health*** *will in fact be determined primarily by the* ***growth in per capita health care costs***"
    - Orszag and Ellis, NEJM, 357:1793, 2007

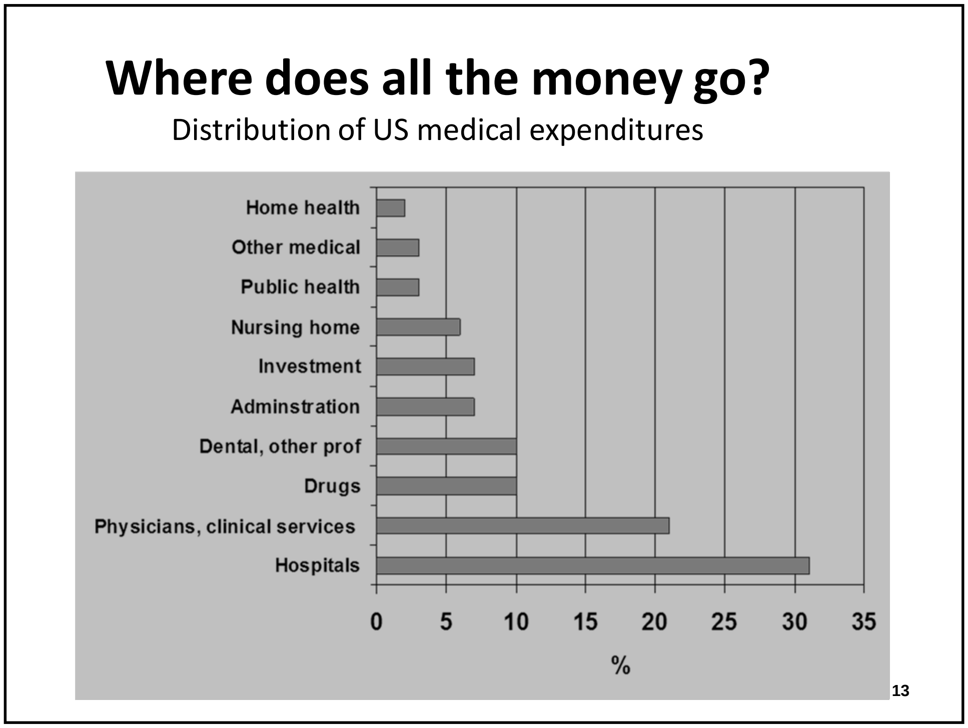
11



Projected Federal Spending for Medicare and Medicaid under Various Assumptions about the Growth Differential between Health Care Costs and per Capita GDP

Orszag and Ellis, NEJM, 357:1793, 2007

12



Where does all the money go?

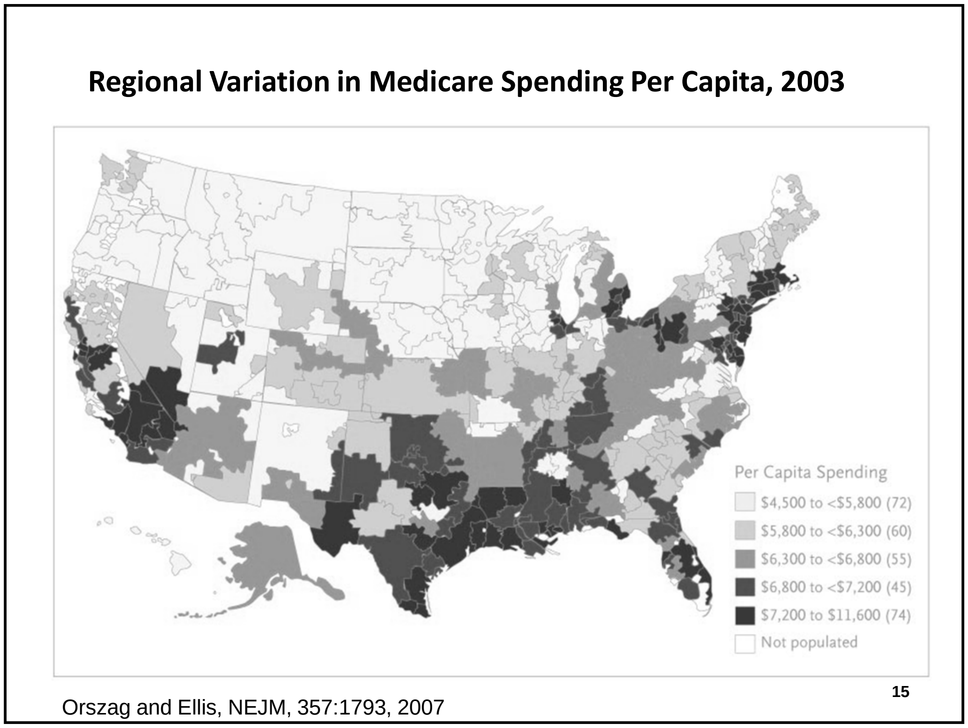
Distribution of US medical expenditures

13

Increased medical costs are…

- Driven primarily by the use of new medical therapies and technologies
  - many of which are not proven to be better or more cost-effective than existing treatments.
- Use of medical services is encouraged by the
  - fee-for-service model (rewards providers for delivering more care e.g., procedures and tests), and
  - lack of incentives for consumers to lessen their demand for services.
- Evidence that higher spending promotes better health outcomes and/or higher quality care is slim to none.

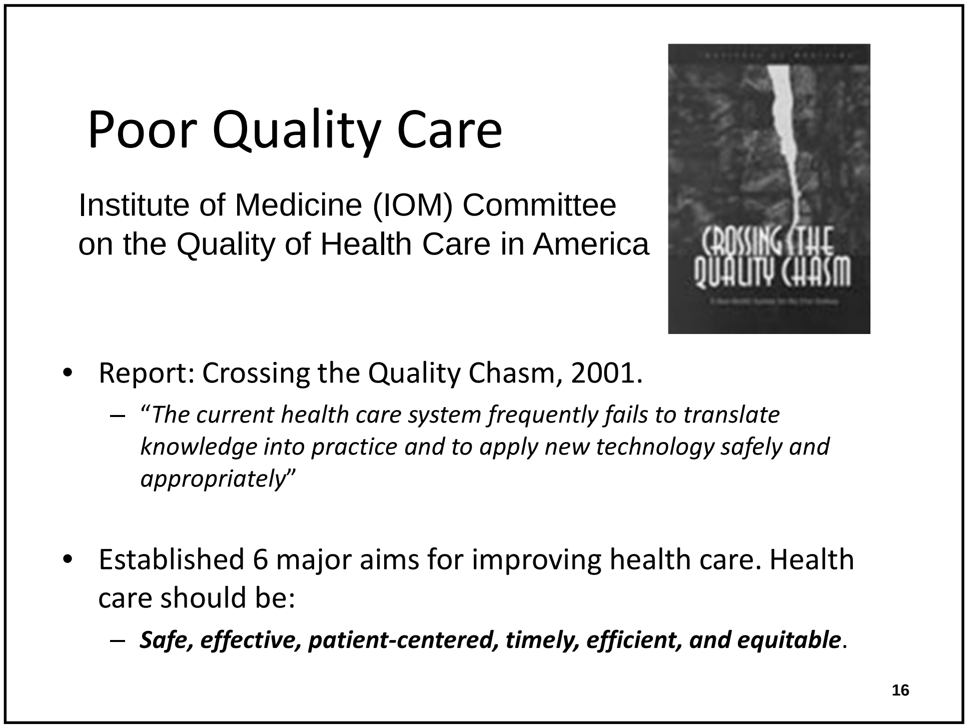
14



Regional Variation in Medicare Spending Per Capita, 2003

Orszag and Ellis, NEJM, 357:1793, 2007

15



Poor Quality Care

Institute of Medicine (IOM) Committee
on the Quality of Health Care in America

CROSSING THE QUALITY CHASM

- Report: Crossing the Quality Chasm, 2001.
  - “*The current health care system frequently fails to translate knowledge into practice and to apply new technology safely and appropriately*”
- Established 6 major aims for improving health care. Health care should be:
  - ***Safe, effective, patient-centered, timely, efficient, and equitable***.

16

Goals of the Epidemiology and Biostatistics Courses

- **Epi-546** (SS 1st Year)
  - To provide a <u>grounding</u> in the principles of clinical epidemiology, and biostatics (**vocabulary, concepts, definitions, applications**) that are fundamental to EBM.
  - 11 lectures, 2 Application sessions

- **Epi-547** (FS 2nd Year)
  - Small group sessions designed to <u>further develop</u> the concepts, definitions and applications of EBM, and to apply them in the evaluation of clinical studies (**critical appraisal**).

17

EBM vocabulary for 21st Century Medicine……

Cost-benefit

Efficacy

OR

RRR

Cost-effective

RR

NNH

95% CI

Effectiveness

Time-to-event analysis

P-value

Intention-to-treat

Risk versus benefit

DB-PC-RCT

Sensitivity

ARR

Meta-analysis

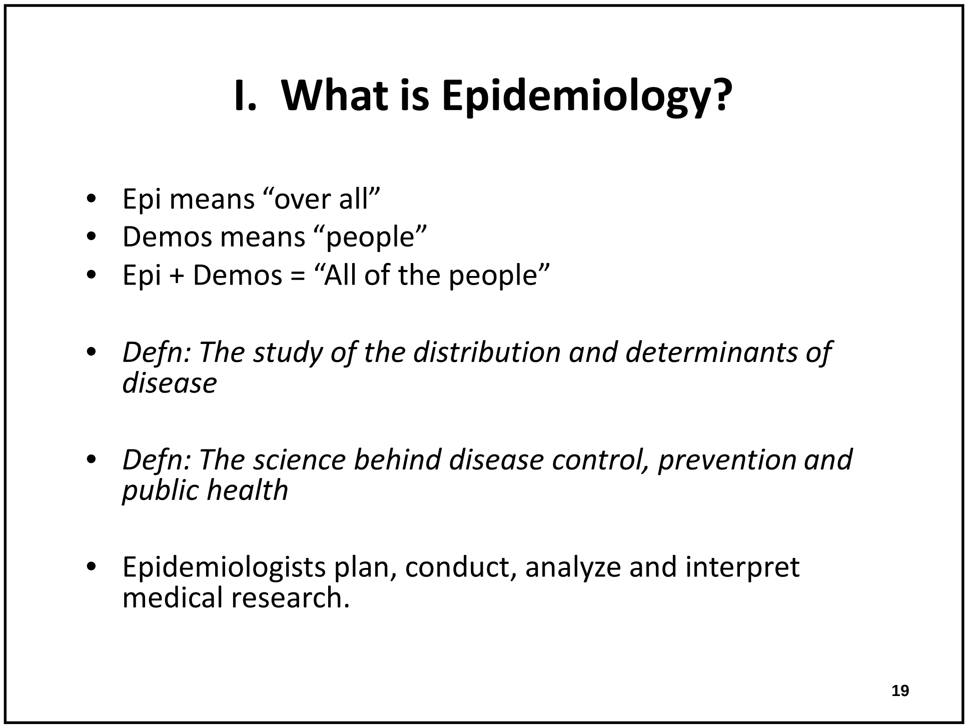
NNT

HR

Likelihood ratio

Population attributable risk

18



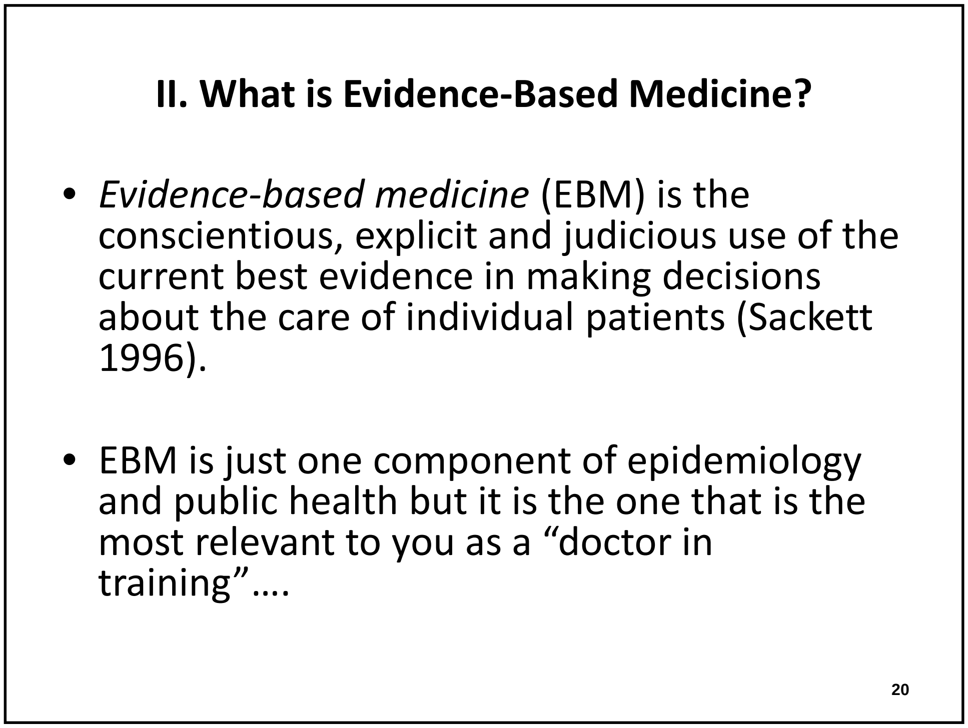
I. What is Epidemiology?

- Epi means “over all”
- Demos means “people”
- Epi + Demos = “All of the people”

- *Defn: The study of the distribution and determinants of disease*

- *Defn: The science behind disease control, prevention and public health*

- Epidemiologists plan, conduct, analyze and interpret medical research.

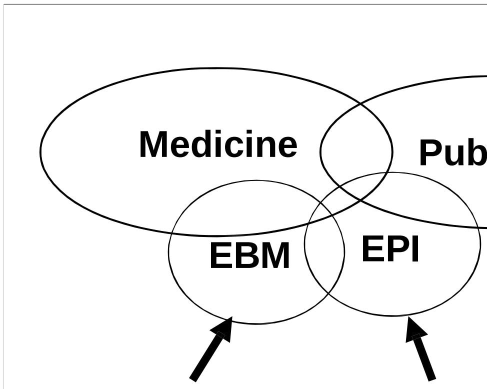


II. What is Evidence-Based Medicine?

- *Evidence-based medicine* (EBM) is the conscientious, explicit and judicious use of the current best evidence in making decisions about the care of individual patients (Sackett 1996).

- EBM is just one component of epidemiology and public health but it is the one that is the most relevant to you as a “doctor in training”….

Relationship between Clinical Medicine, Public Health, EBM and Epidemiology…….



MDs trained in EBM

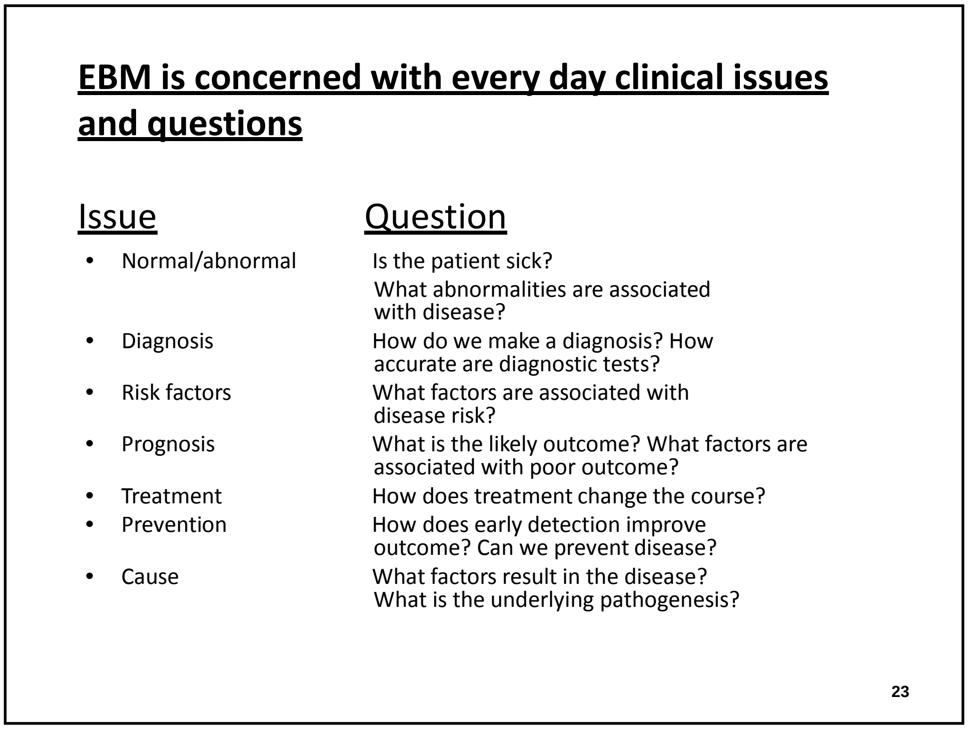
Health professionals with PhD, MS, or MPH

21

EBM - Important Concepts

- Synthesis of individual clinical expertise and external evidence from systematic research.
- Stresses expertise in information gathering, synthesis and incorporation.
- De-emphasizes memorization.
- Rebellious disregard for authoritarian “expert opinion”.
- Relies heavily on medical informatics and on-line resources (e.g., PubMed, ACP Journal Club, Cochrane Database of Systematic Reviews)

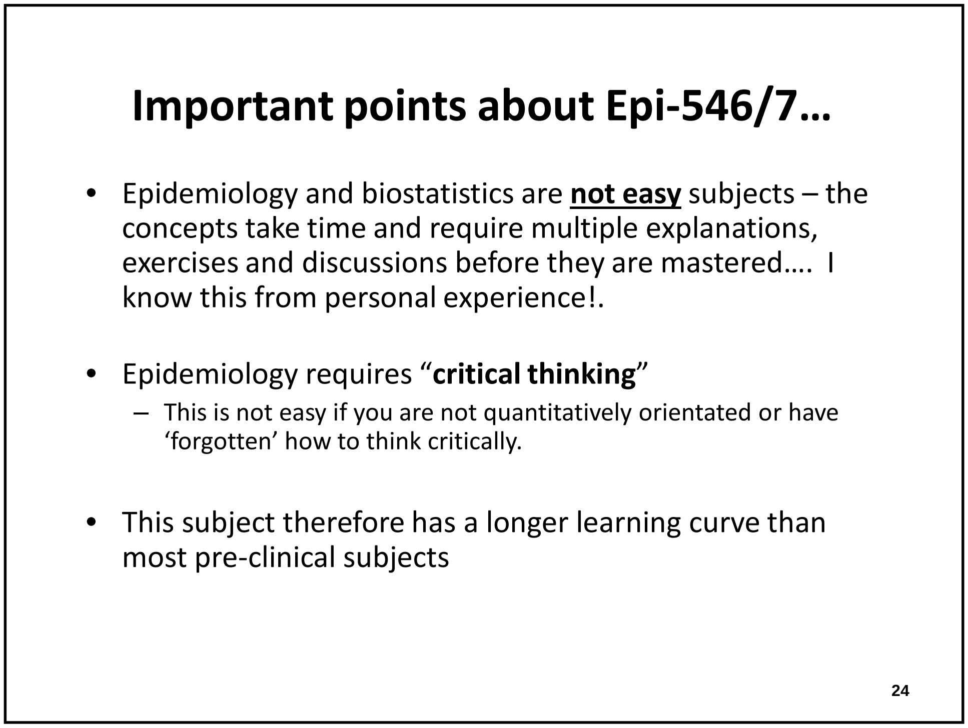
22



EBM is concerned with every day clinical issues and questions

Issue	Question
• Normal/abnormal	Is the patient sick? What abnormalities are associated with disease?
• Diagnosis	How do we make a diagnosis? How accurate are diagnostic tests?
• Risk factors	What factors are associated with disease risk?
• Prognosis	What is the likely outcome? What factors are associated with poor outcome?
• Treatment	How does treatment change the course?
• Prevention	How does early detection improve outcome? Can we prevent disease?
• Cause	What factors result in the disease? What is the underlying pathogenesis?

23

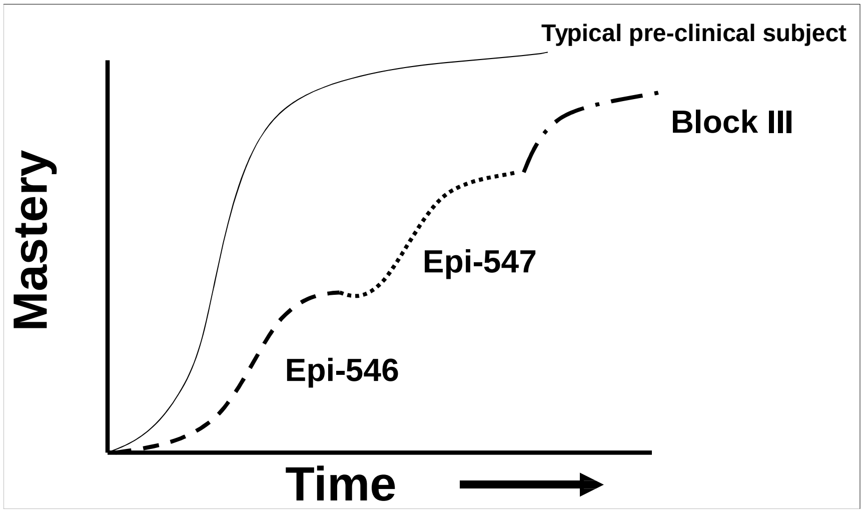


Important points about Epi-546/7…

- Epidemiology and biostatistics are **not easy** subjects – the concepts take time and require multiple explanations, exercises and discussions before they are mastered…. I know this from personal experience!.
- Epidemiology requires "**critical thinking**"
  - This is not easy if you are not quantitatively orientated or have 'forgotten' how to think critically.
- This subject therefore has a longer learning curve than most pre-clinical subjects

24

Epidemiology has a prolonged learning curve



Things about Epi-546 that you will come to appreciate …

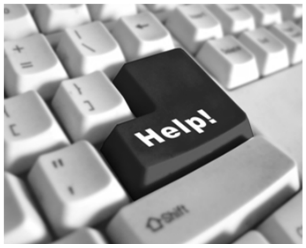
- This seems like a lot of work for a 1 credit course
    - We agree. Think of it as a 2 credit course if that helps.
- This seems like more information than is needed for Boards?
    - We agree. We are not that interested in boards. We are focused on the 30+ years that come afterwards.
- I have an MPH. Why do I have to take this course?
    - This is clinical not public health epidemiology – its different.
- 8 am lectures in January?
    - Tell us about it….

My point about medical boards….

Boards are like <u>potty training,</u> they represent an important and necessary step, but once achieved it is important to move on and accomplish other things in life.

27

There is lots of help….



- On-line practice questions on D2L
- Two '*Application Sessions*' designed to reinforce concepts
- Four practice exams (2 mid-term, 2 final) on D2L
- Directed Study Groups (DSG) - supplemental help from PhD epidemiology graduate students (if needed)
- Discussion board on D2L for posting questions

28

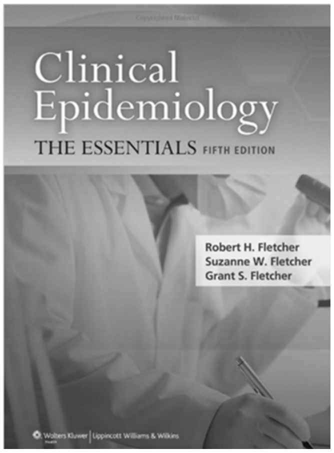
The Epi-546 Course

- There is a course pack –get it and read it!
  - Read the course polices carefully!
- ALL THE CONTENTS OF THE COURSE PACK ARE FAIR GAME FOR THE EXAMS
- All materials are also under the “Lessons” folder on D2L.
- Within each folder you will generally find:
  - .pdf of the core PowerPoint lectures slides
  - .pdf course notes
  - Practice questions
  - Pre-recorded lecture (Lecture 2)
- All “live lectures” will be recorded and put on the Mediasite.

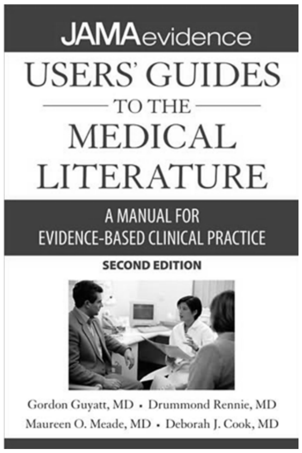
29

Text Books

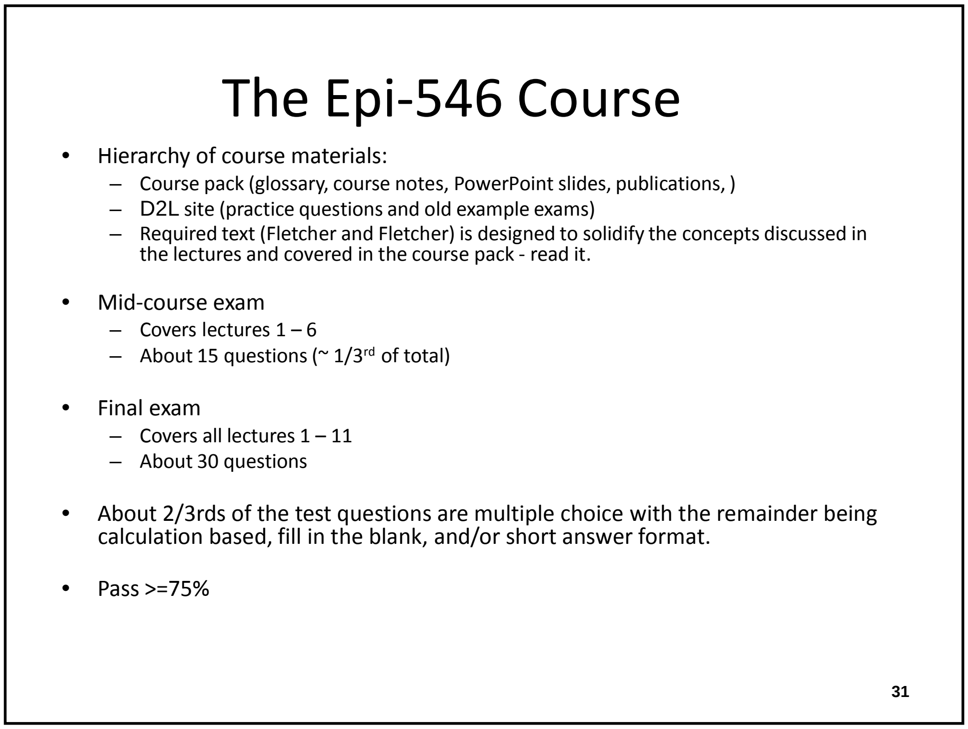
Required


Clinical
Epidemiology
THE ESSENTIALS FIFTH EDITION
Robert H. Fletcher
Suzanne W. Fletcher
Grant S. Fletcher
Wolters Kluwer | Lippincott Williams & Wilkins

Optional

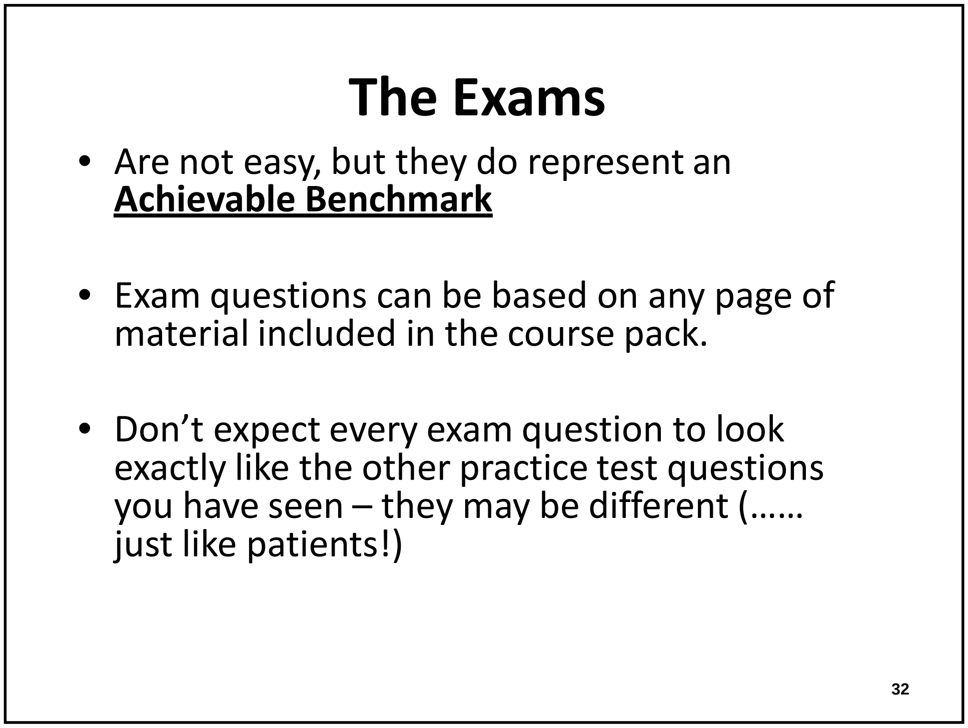

JAMAevidence
USERS' GUIDES
TO THE
MEDICAL
LITERATURE
A MANUAL FOR
EVIDENCE-BASED CLINICAL PRACTICE
SECOND EDITION
Gordon Guyatt, MD • Drummond Rennie, MD
Maureen O. Meade, MD • Deborah J. Cook, MD

30



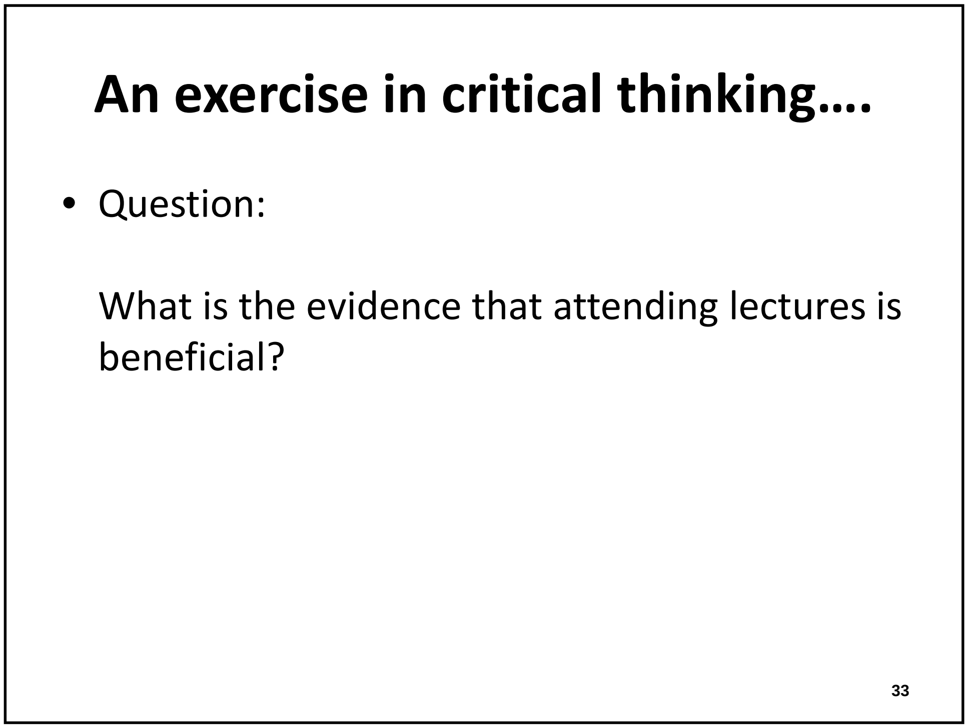
The Epi-546 Course

- Hierarchy of course materials:
  - Course pack (glossary, course notes, PowerPoint slides, publications,)
  - D2L site (practice questions and old example exams)
  - Required text (Fletcher and Fletcher) is designed to solidify the concepts discussed in the lectures and covered in the course pack - read it.
- Mid-course exam
  - Covers lectures 1 – 6
  - About 15 questions (~ 1/3rd of total)
- Final exam
  - Covers all lectures 1 – 11
  - About 30 questions
- About 2/3rds of the test questions are multiple choice with the remainder being calculation based, fill in the blank, and/or short answer format.
- Pass >=75%



The Exams

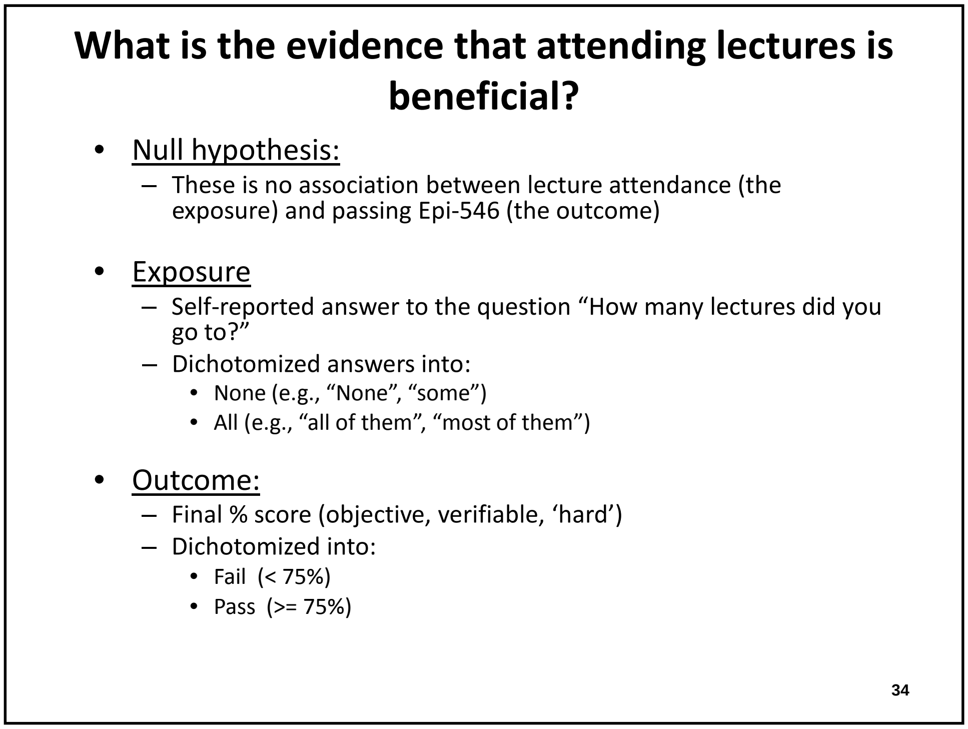
- Are not easy, but they do represent an **<u>Achievable Benchmark</u>**
- Exam questions can be based on any page of material included in the course pack.
- Don’t expect every exam question to look exactly like the other practice test questions you have seen – they may be different (…… just like patients!)



An exercise in critical thinking….

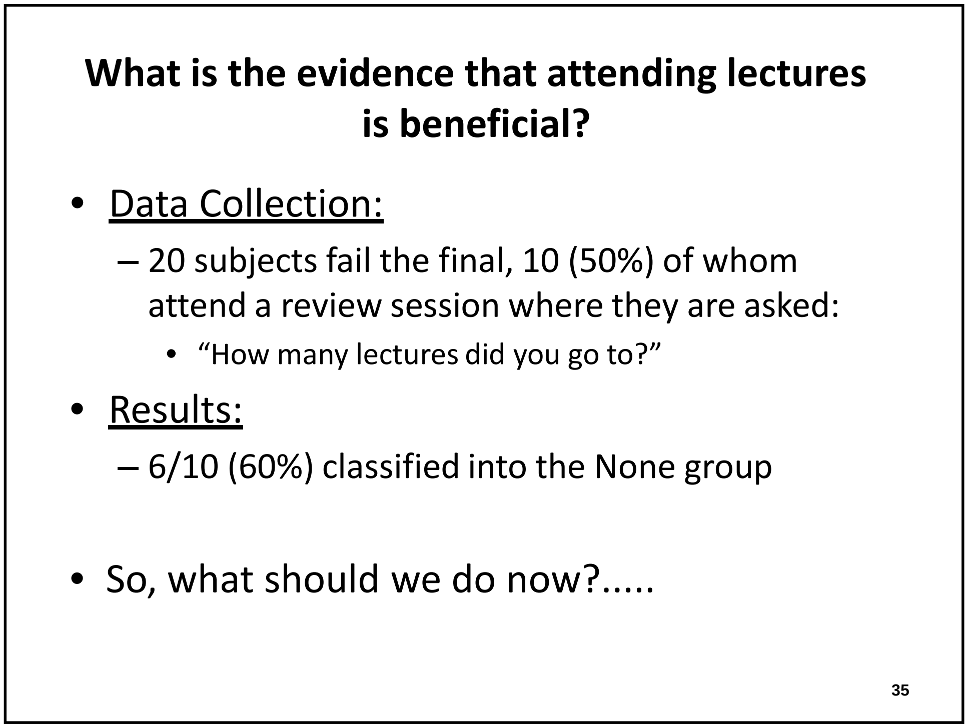
- Question:

  What is the evidence that attending lectures is beneficial?



What is the evidence that attending lectures is beneficial?

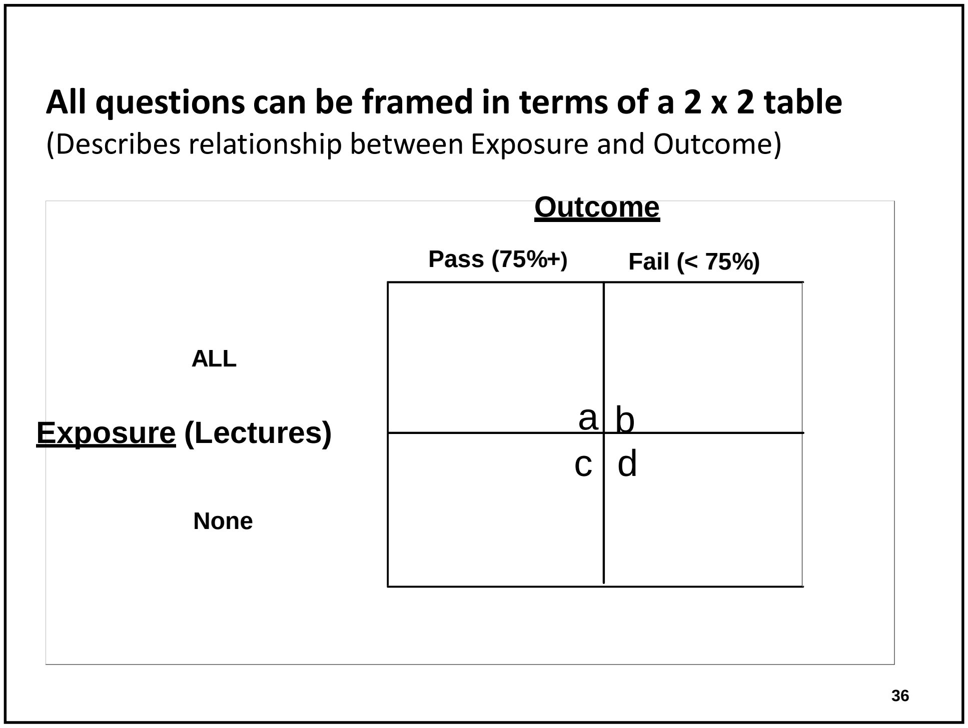
- Null hypothesis:
  - These is no association between lecture attendance (the exposure) and passing Epi-546 (the outcome)
- Exposure
  - Self-reported answer to the question “How many lectures did you go to?”
  - Dichotomized answers into:
    - None (e.g., “None”, “some”)
    - All (e.g., “all of them”, “most of them”)
- Outcome:
  - Final % score (objective, verifiable, ‘hard’)
  - Dichotomized into:
    - Fail (< 75%)
    - Pass (>= 75%)



What is the evidence that attending lectures is beneficial?

- <u>Data Collection:</u>
  - 20 subjects fail the final, 10 (50%) of whom attend a review session where they are asked:
    - “How many lectures did you go to?”
- <u>Results:</u>
  - 6/10 (60%) classified into the None group
- So, what should we do now?.....

35



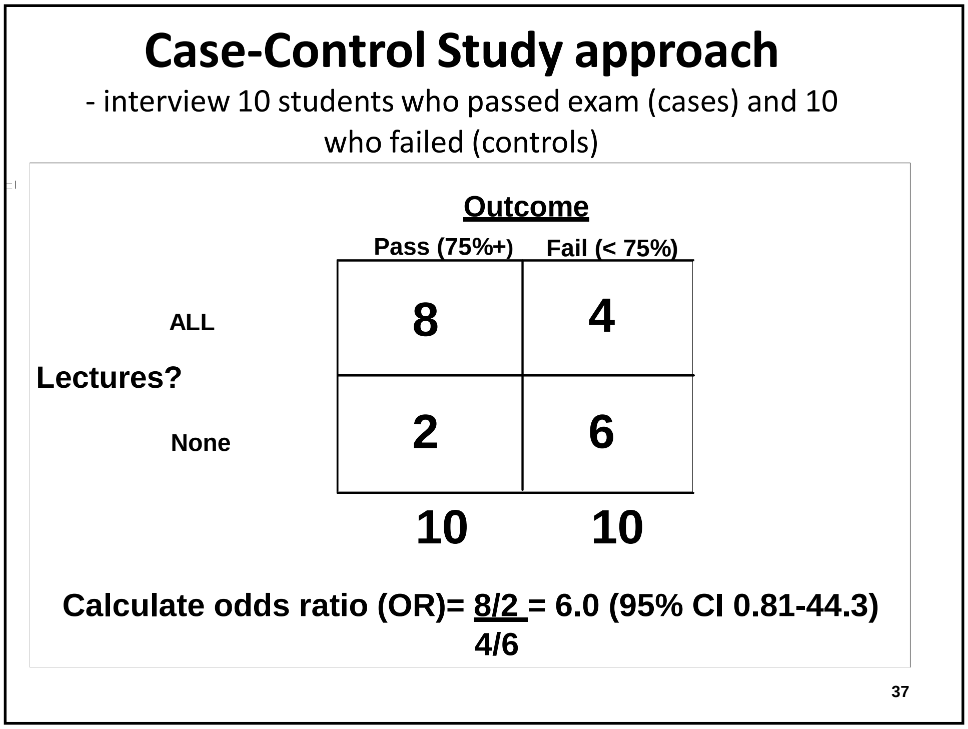
All questions can be framed in terms of a 2 x 2 table

(Describes relationship between Exposure and Outcome)

<u>Outcome</u>

		Pass (75%+)	Fail (< 75%)
<u>Exposure</u> (Lectures)	**ALL**	a	b
	None	c	d

36



Case-Control Study approach

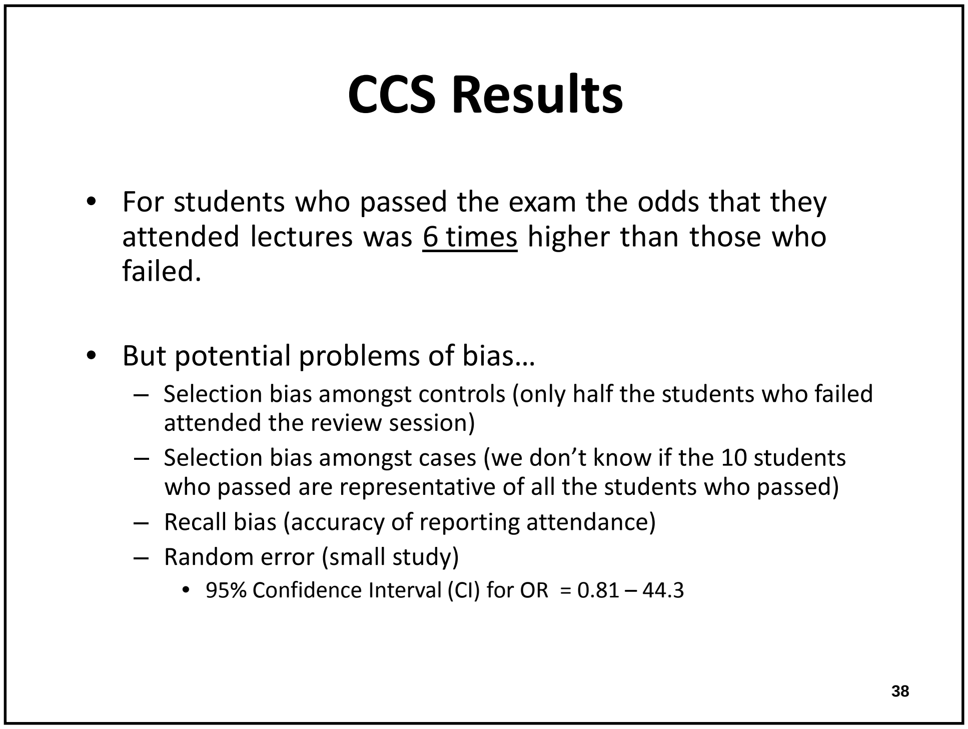
- interview 10 students who passed exam (cases) and 10 who failed (controls)

Outcome

		Pass (75%+)	Fail (< 75%)
Lectures?	ALL	8	4
	None	2	6
		10	10

Calculate odds ratio (OR)= $\frac{8/2}{4/6}$ = 6.0 (95% CI 0.81-44.3)

37



CCS Results

- For students who passed the exam the odds that they attended lectures was <u>6 times</u> higher than those who failed.
- But potential problems of bias…
  - Selection bias amongst controls (only half the students who failed attended the review session)
  - Selection bias amongst cases (we don't know if the 10 students who passed are representative of all the students who passed)
  - Recall bias (accuracy of reporting attendance)
  - Random error (small study)
    - 95% Confidence Interval (CI) for OR = 0.81 – 44.3

38

Alternative approach – a cohort study

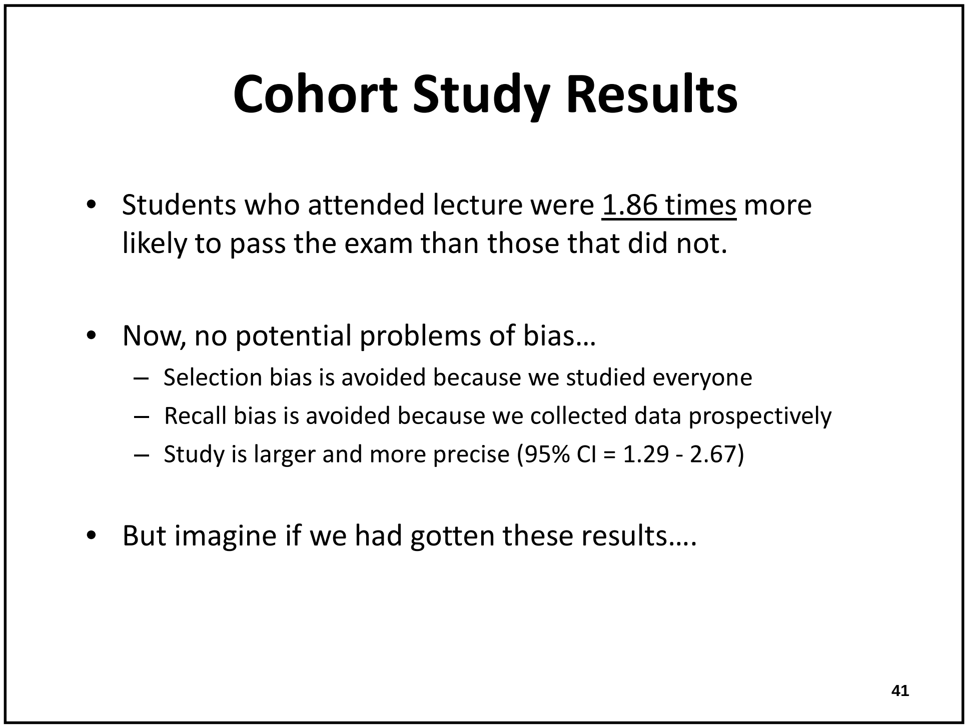
- Collect data prospectively on attendance before the final exam
- But how to collect such data?... Ideas?.....
- Imagine that 70% of the class were found to meet the definition of “ALL”.
- Now correlate this with the exam results….

Cohort Study approach

- follow all 100 students, 70% attend lectures

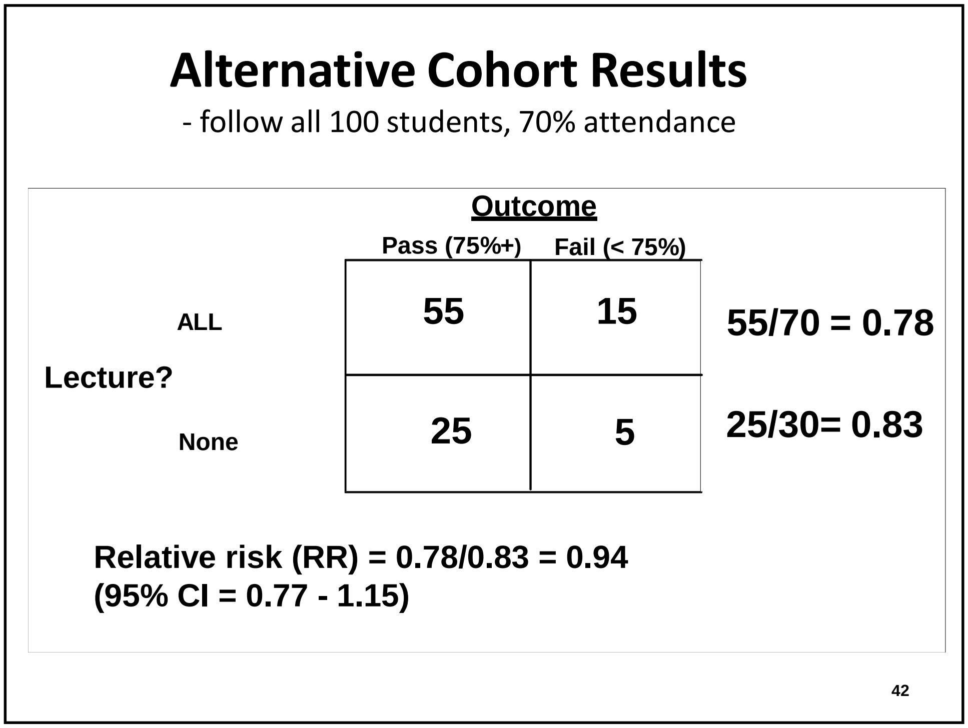
		Outcome		
		Pass (75%+)	Fail (< 75%)	
Lectures?	ALL	65	5	65/70 = 0.92
	None	15	15	15/30 = 0.50

Calculate relative risk (RR) = 0.92/0.5 = 1.86



Cohort Study Results

- Students who attended lecture were 1.86 times more likely to pass the exam than those that did not.
- Now, no potential problems of bias…
  - Selection bias is avoided because we studied everyone
  - Recall bias is avoided because we collected data prospectively
  - Study is larger and more precise (95% CI = 1.29 - 2.67)
- But imagine if we had gotten these results….



Alternative Cohort Results

- follow all 100 students, 70% attendance

		Outcome		
		Pass (75%+)	**Fail (< 75%)**	
Lecture?	**ALL**	**55**	**15**	**55/70 = 0.78**
	None	**25**	**5**	**25/30= 0.83**

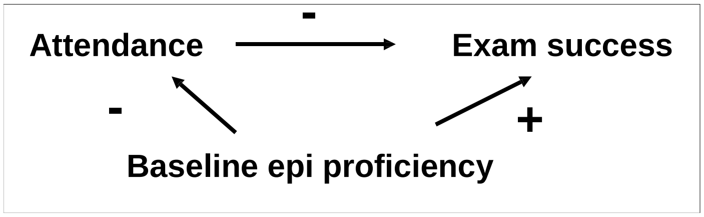
Relative risk (RR) = 0.78/0.83 = 0.94
(95% CI = 0.77 - 1.15)

Alternative Cohort Study Results

- Students who attended lecture were <u>0.94 times</u> **less** likely to pass the exam than those that did not.
- Why could this be a plausible finding?…….

Because of <u>confounding…</u>…..

- Students who chose not to attend had greater baseline proficiency in epidemiology (prior education? or maybe higher IQ)

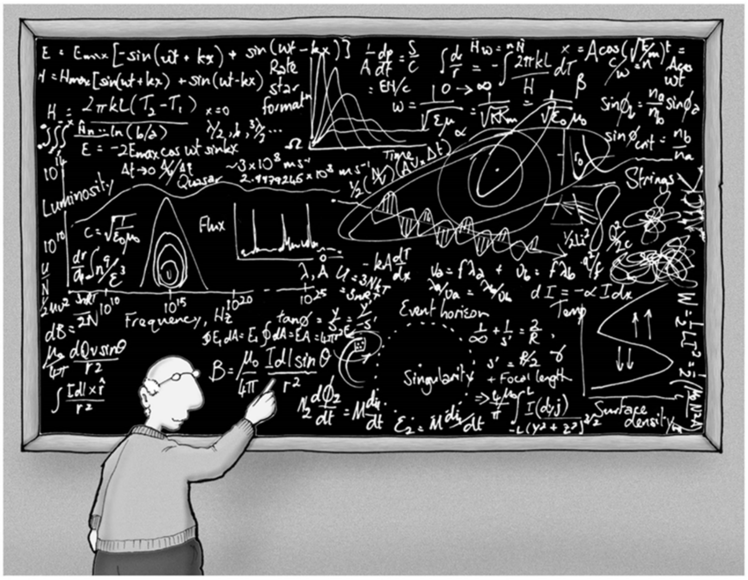


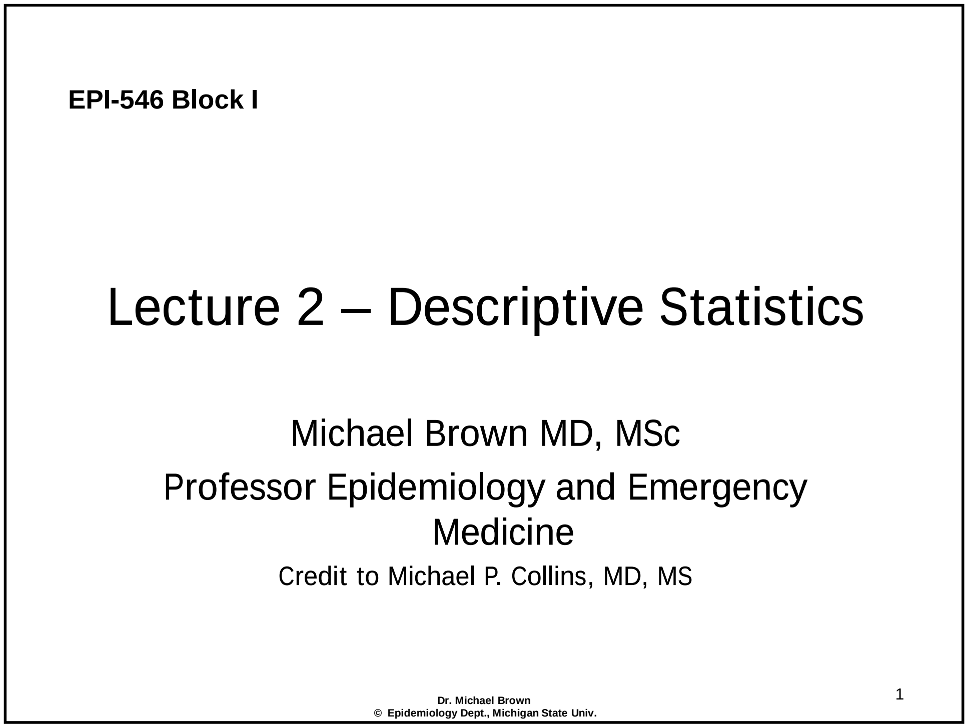
Attendance → **Exam success** (-)

Baseline epi proficiency → **Attendance** (-)

Baseline epi proficiency → **Exam success** (+)

What is the fix for this problem??


Rate
star
formation
Luminosity
Quasar
Flux
Frequency, Hz
Time
Strings
Event horison
Temp
Singularity
+ Focal length
Surface
density



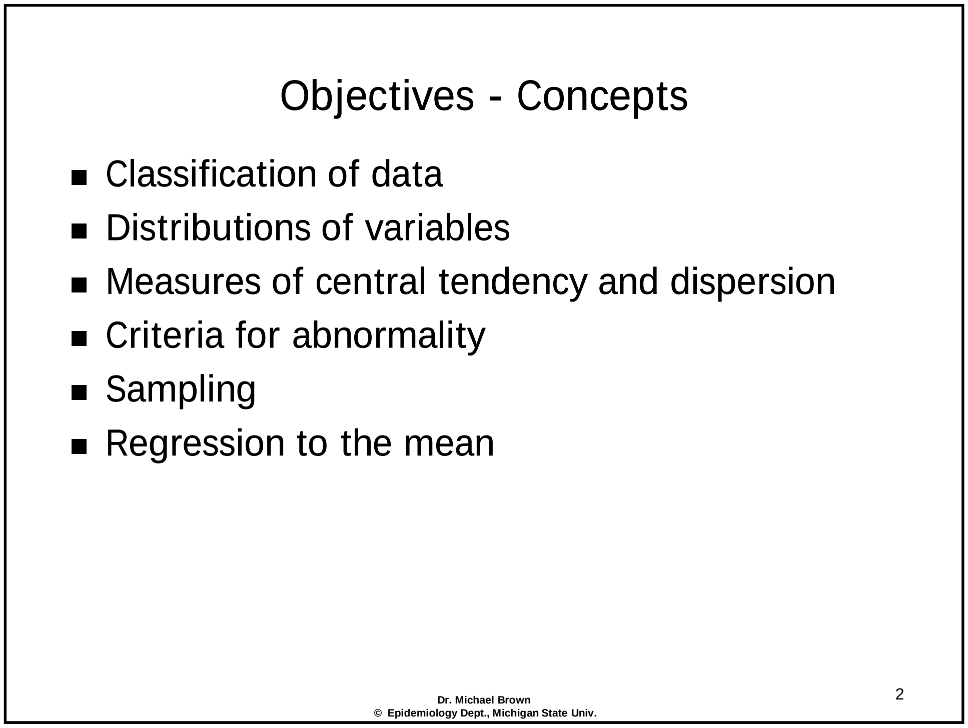
EPI-546 Block I

Lecture 2 – Descriptive Statistics

Michael Brown MD, MSc

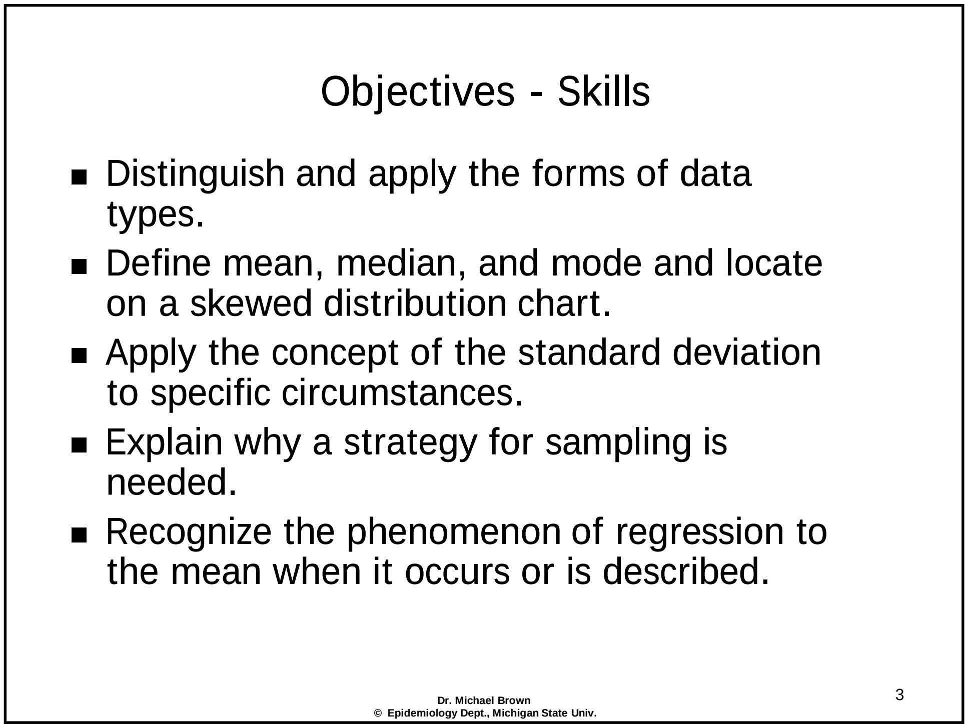
Professor Epidemiology and Emergency Medicine

Credit to Michael P. Collins, MD, MS



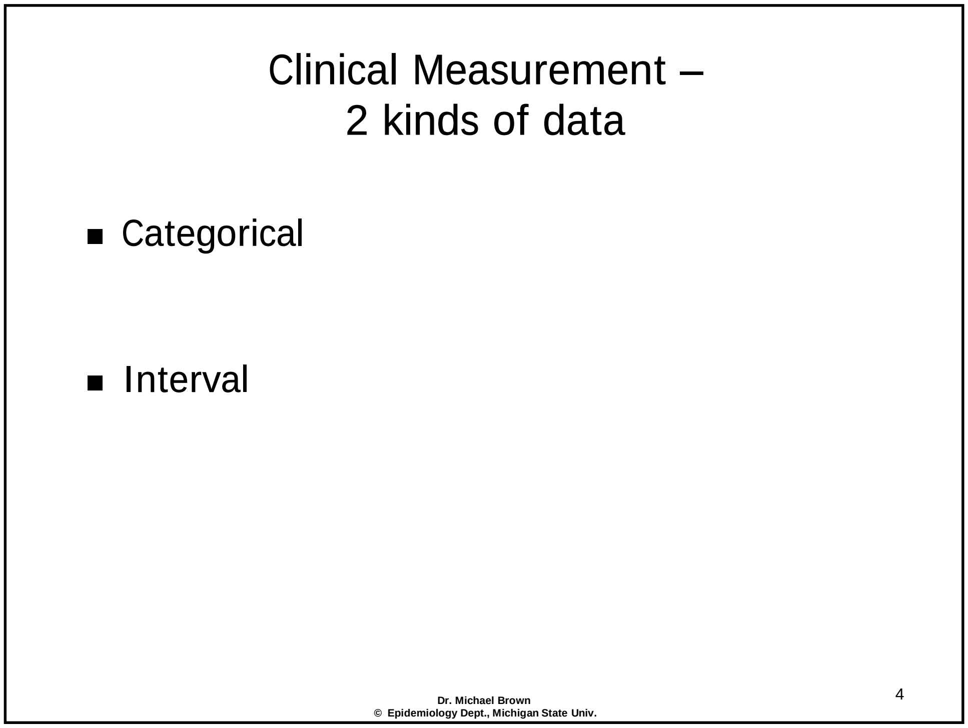
Objectives - Concepts

- Classification of data
- Distributions of variables
- Measures of central tendency and dispersion
- Criteria for abnormality
- Sampling
- Regression to the mean



Objectives - Skills

- Distinguish and apply the forms of data types.
- Define mean, median, and mode and locate on a skewed distribution chart.
- Apply the concept of the standard deviation to specific circumstances.
- Explain why a strategy for sampling is needed.
- Recognize the phenomenon of regression to the mean when it occurs or is described.



Clinical Measurement – 2 kinds of data

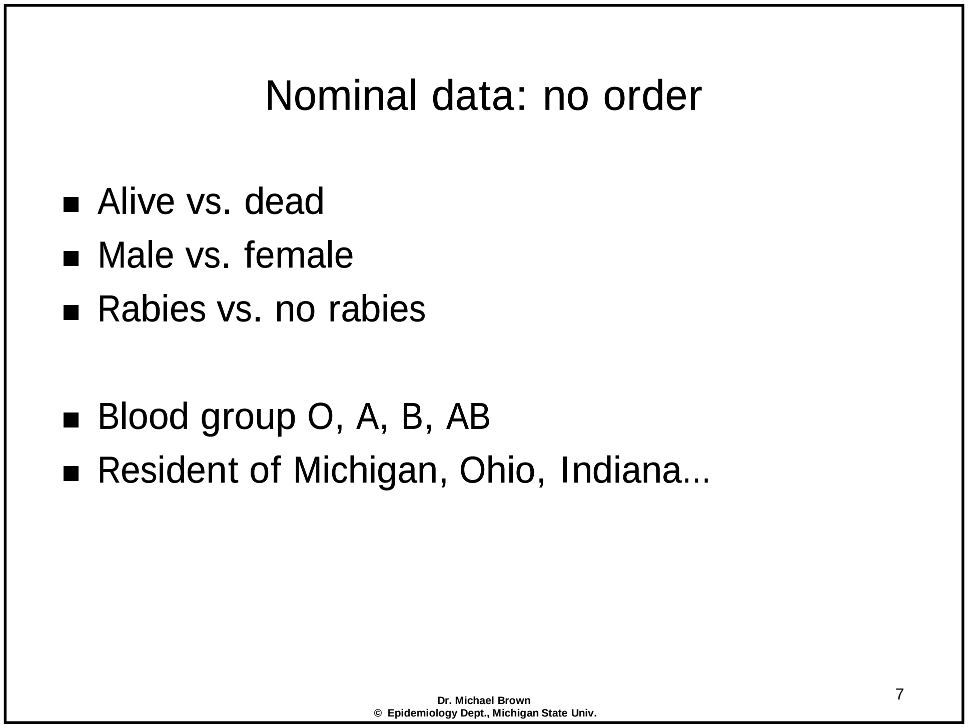
- Categorical
- Interval

Distinction -

Interval = “the interval between successive values is equal, throughout the scale”

Clinical Measurement – subtypes of data

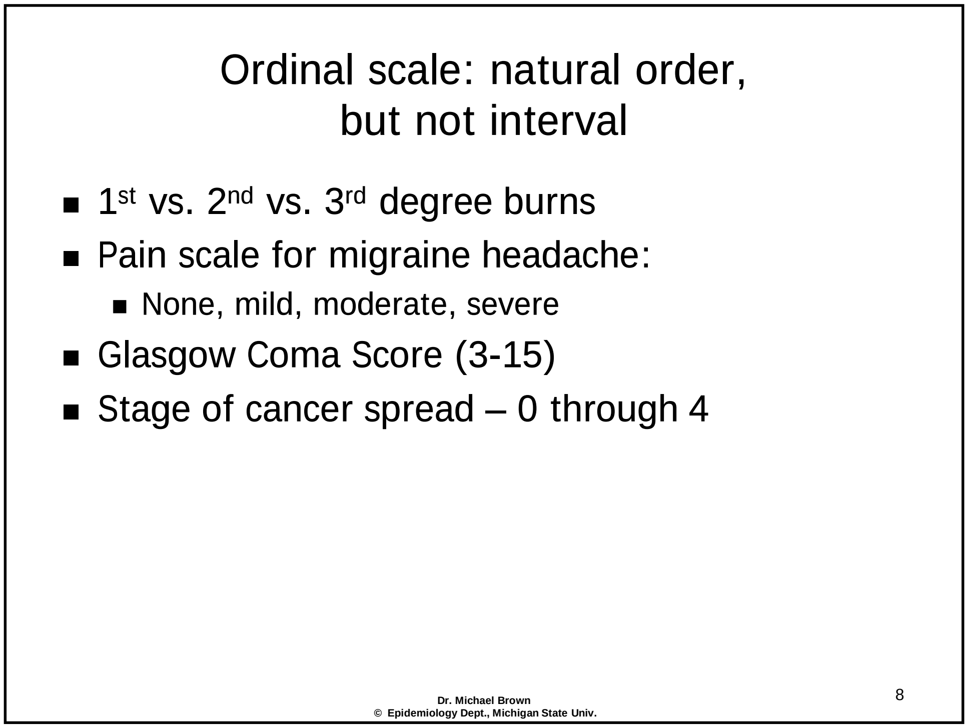
- Categorical
  - Nominal
  - Ordinal
- Interval
  - Discrete
  - Continuous



Nominal data: no order

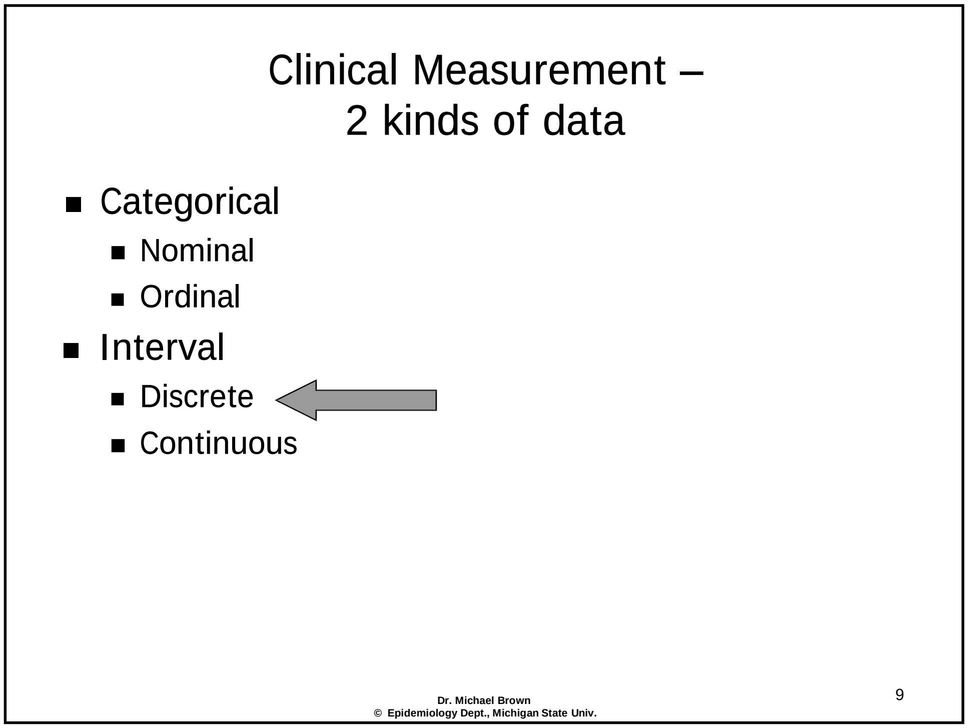
- Alive vs. dead
- Male vs. female
- Rabies vs. no rabies

- Blood group O, A, B, AB
- Resident of Michigan, Ohio, Indiana...



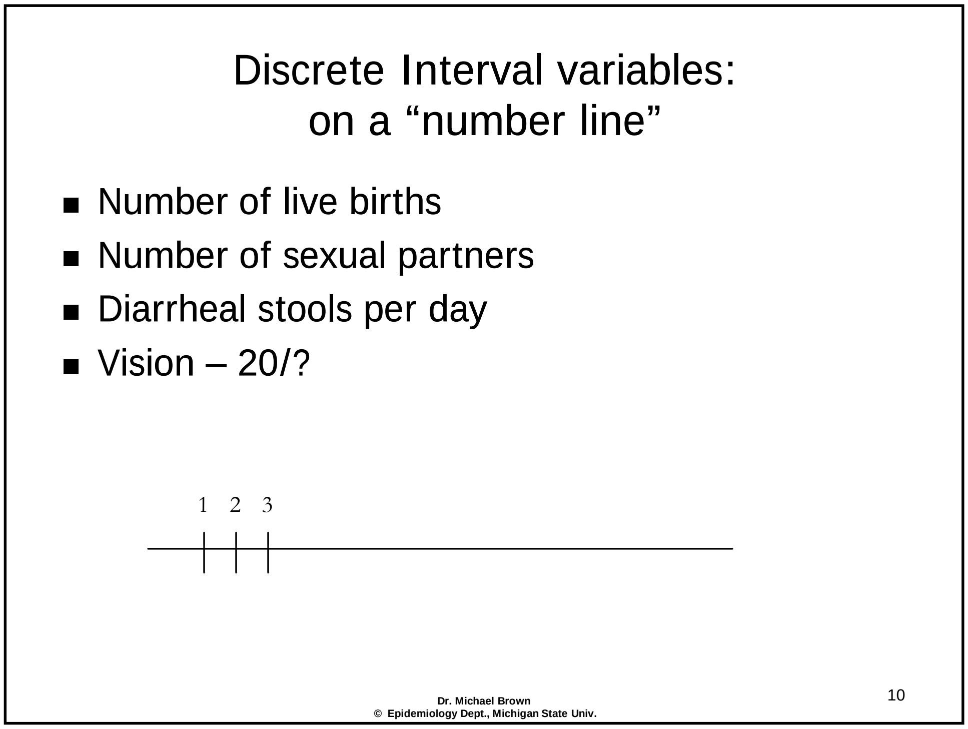
Ordinal scale: natural order, but not interval

- 1st vs. 2nd vs. 3rd degree burns
- Pain scale for migraine headache:
  - None, mild, moderate, severe
- Glasgow Coma Score (3-15)
- Stage of cancer spread – 0 through 4



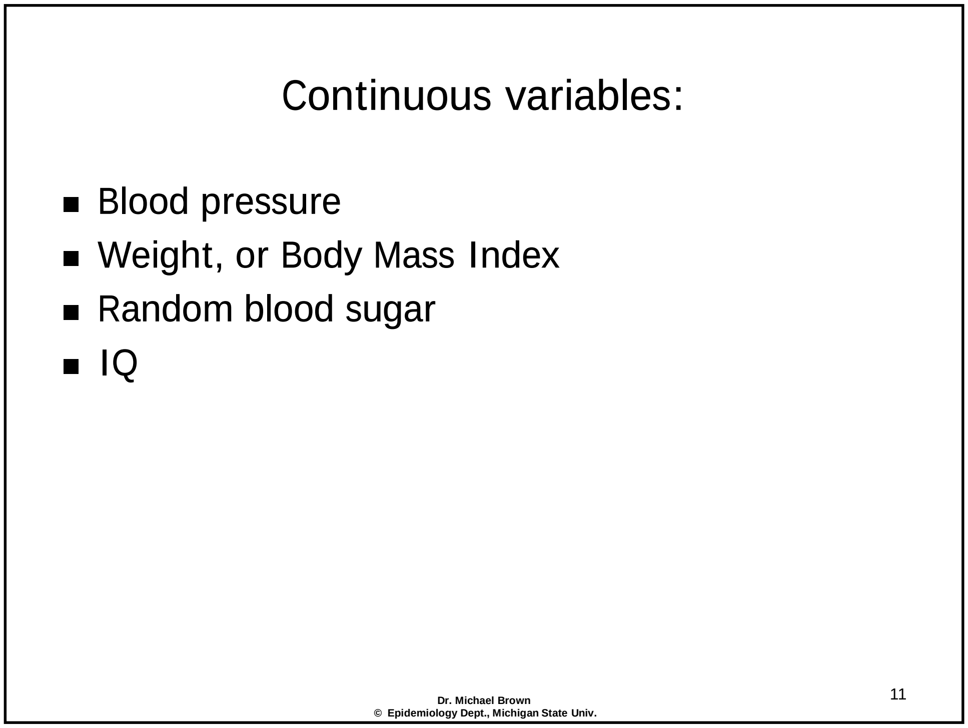
Clinical Measurement – 2 kinds of data

- Categorical
  - Nominal
  - Ordinal
- Interval
  - Discrete
  - Continuous



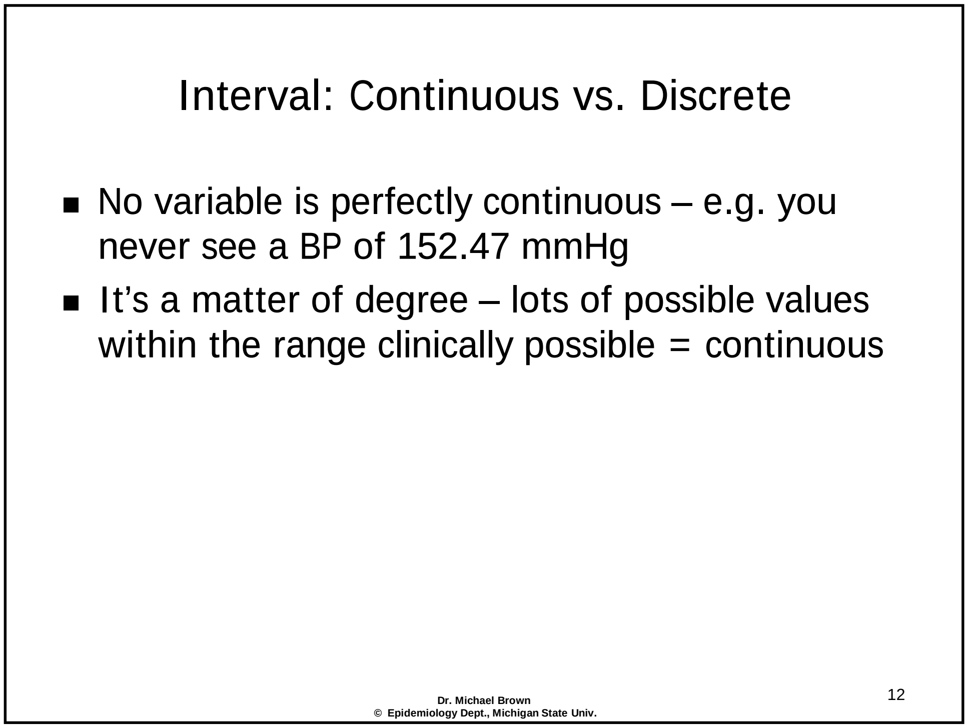
Discrete Interval variables: on a “number line”

- Number of live births
- Number of sexual partners
- Diarrheal stools per day
- Vision – 20/?



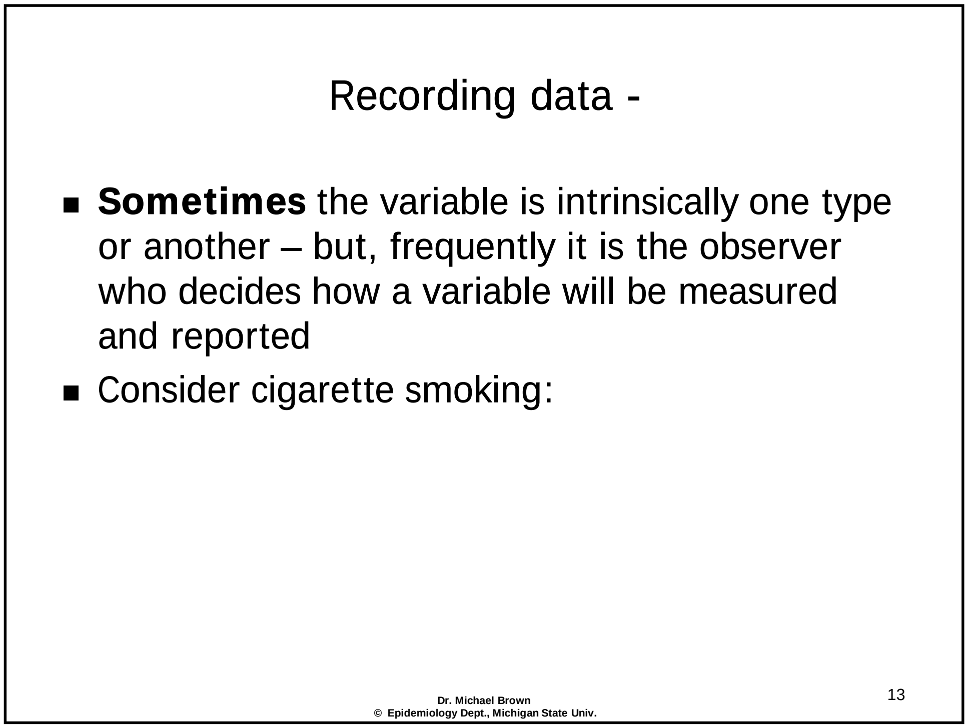
Continuous variables:

- Blood pressure
- Weight, or Body Mass Index
- Random blood sugar
- IQ



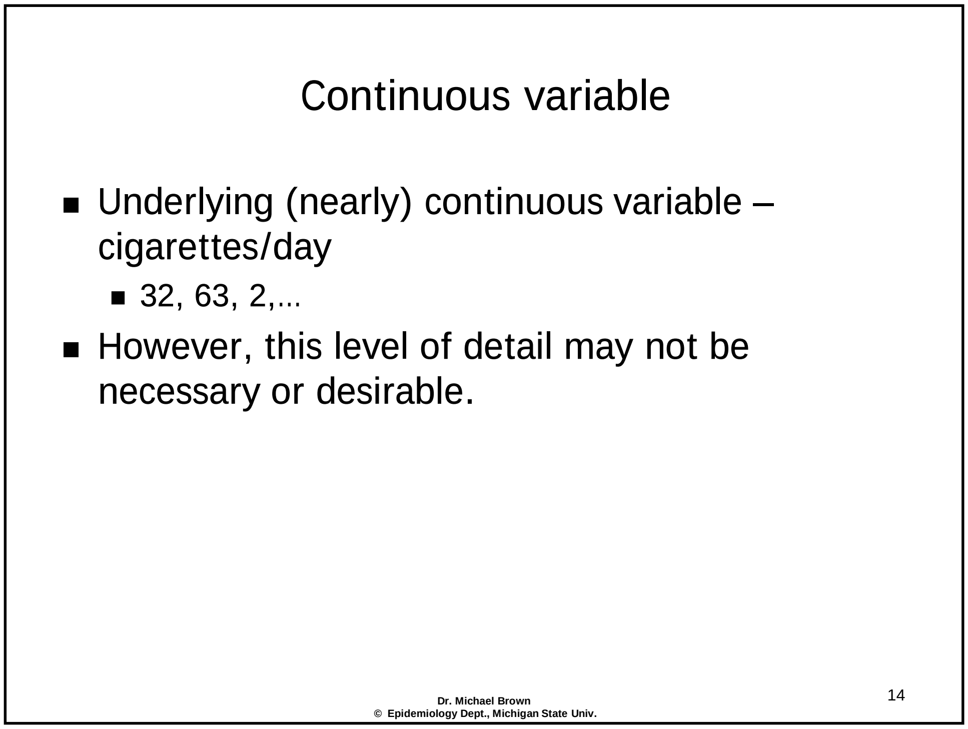
Interval: Continuous vs. Discrete

- No variable is perfectly continuous – e.g. you never see a BP of 152.47 mmHg
- It’s a matter of degree – lots of possible values within the range clinically possible = continuous



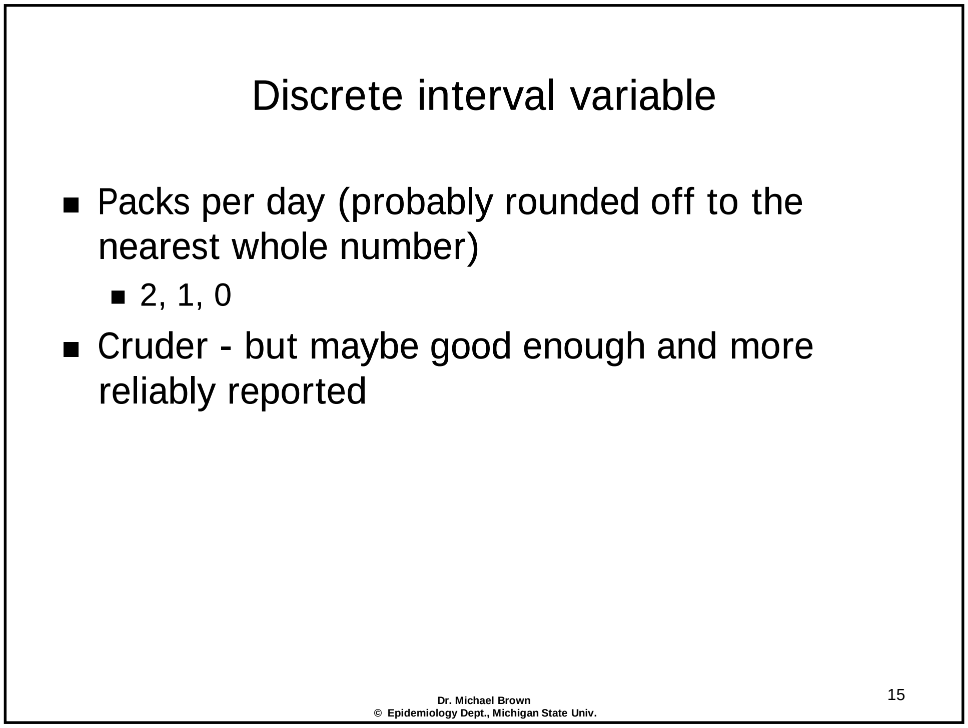
Recording data -

- **Sometimes** the variable is intrinsically one type or another – but, frequently it is the observer who decides how a variable will be measured and reported
- Consider cigarette smoking:



Continuous variable

- Underlying (nearly) continuous variable – cigarettes/day
  - 32, 63, 2,...
- However, this level of detail may not be necessary or desirable.

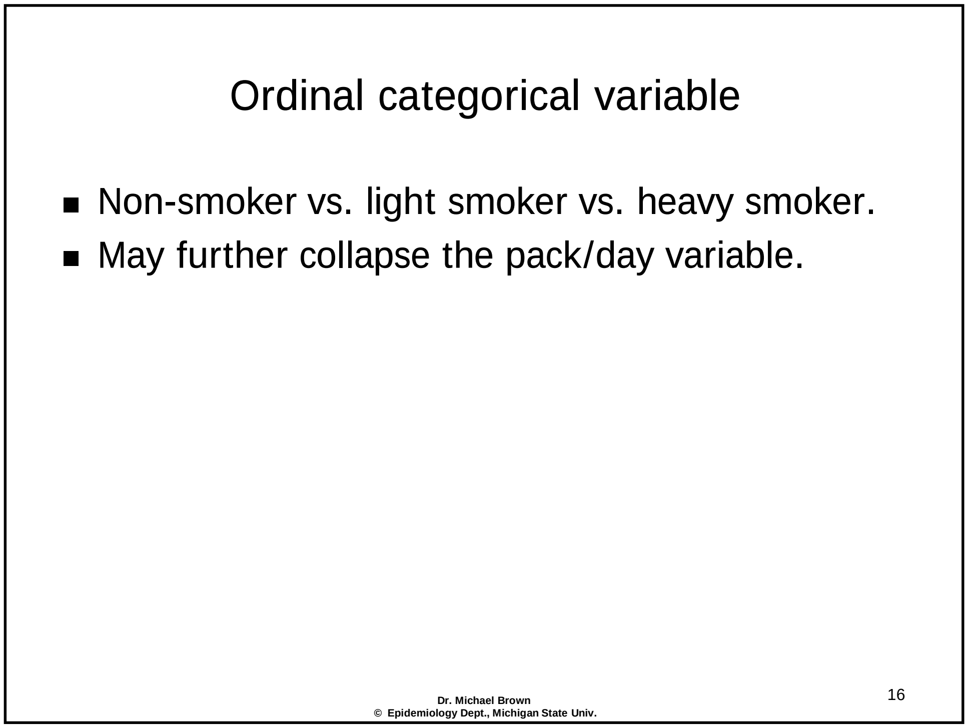


Discrete interval variable

- Packs per day (probably rounded off to the nearest whole number)
  - 2, 1, 0
- Cruder - but maybe good enough and more reliably reported

Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.

15

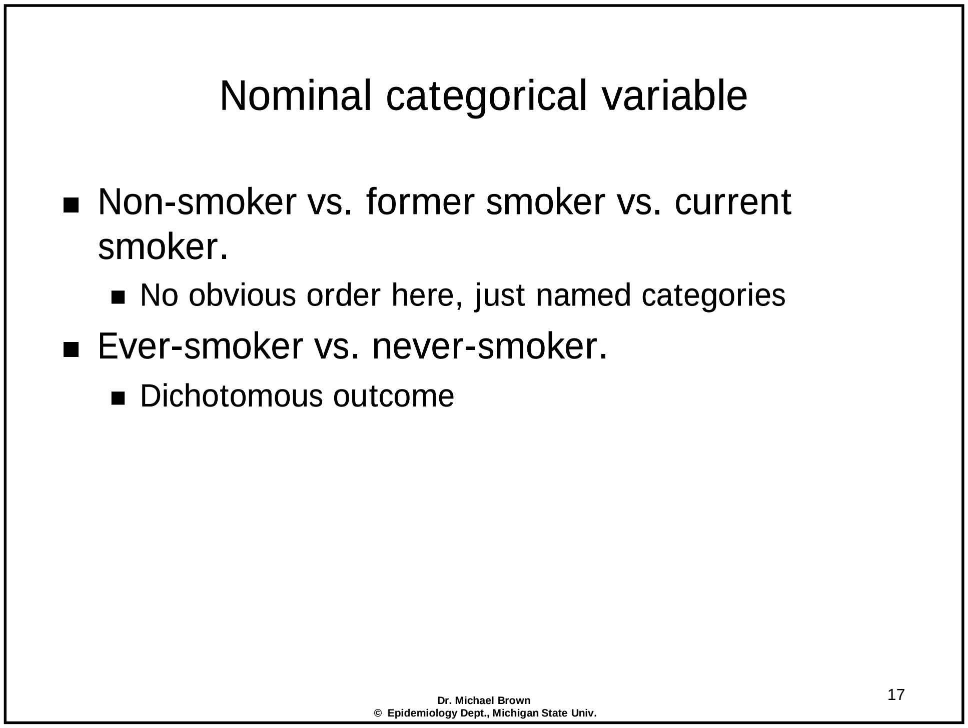


Ordinal categorical variable

- Non-smoker vs. light smoker vs. heavy smoker.
- May further collapse the pack/day variable.

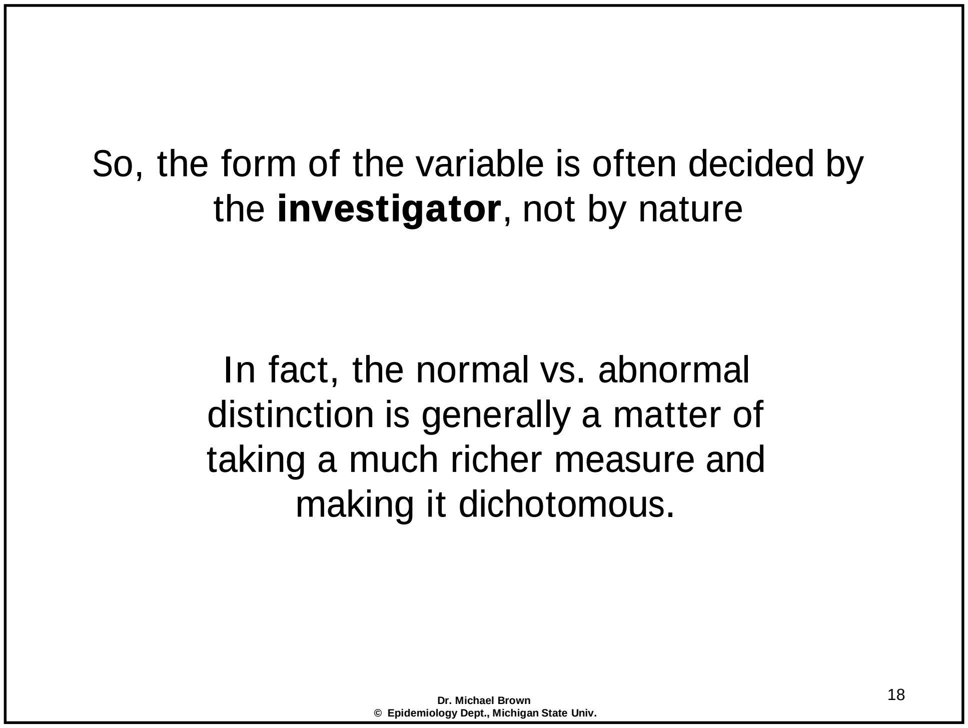
Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.

16



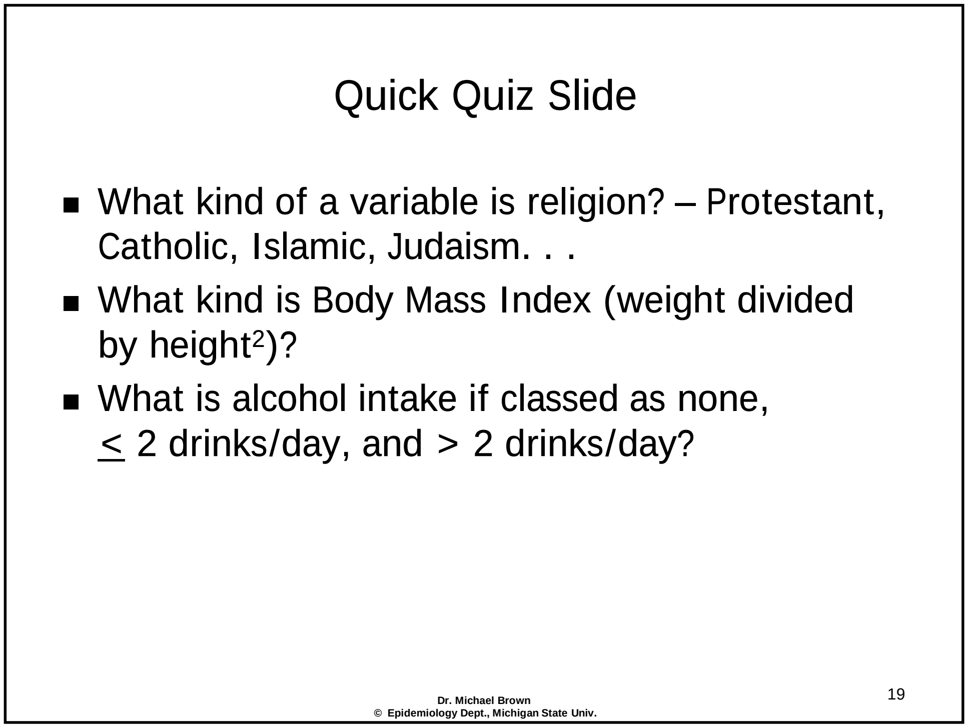
Nominal categorical variable

- Non-smoker vs. former smoker vs. current smoker.
  - No obvious order here, just named categories
- Ever-smoker vs. never-smoker.
  - Dichotomous outcome



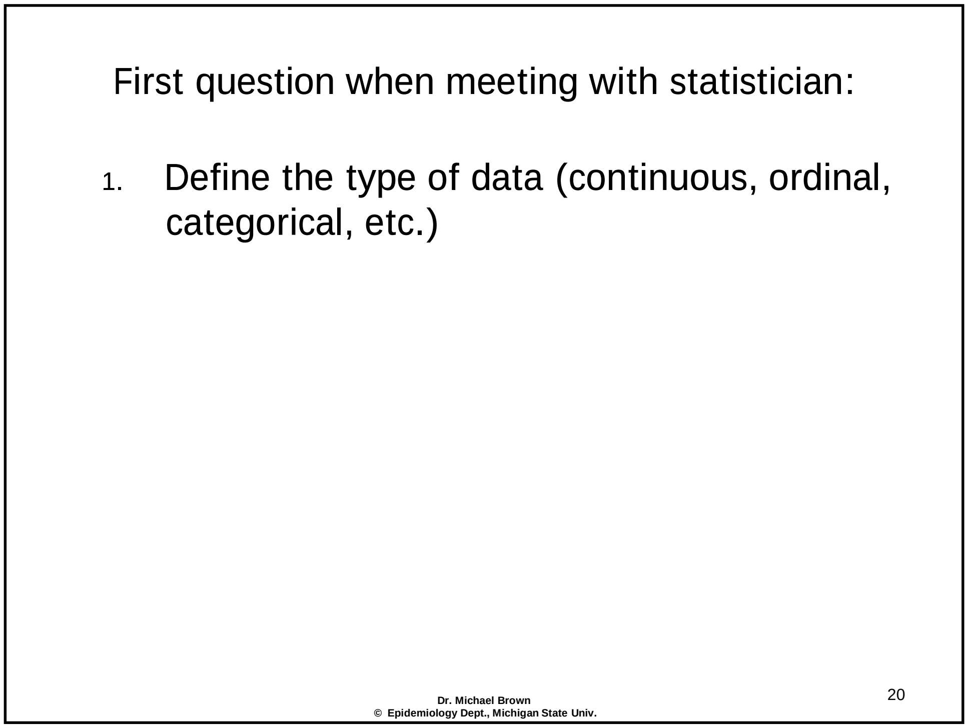
So, the form of the variable is often decided by the **investigator**, not by nature

In fact, the normal vs. abnormal distinction is generally a matter of taking a much richer measure and making it dichotomous.



Quick Quiz Slide

- What kind of a variable is religion? – Protestant, Catholic, Islamic, Judaism. . .
- What kind is Body Mass Index (weight divided by height^2)?
- What is alcohol intake if classed as none, $\leq$ 2 drinks/day, and > 2 drinks/day?



First question when meeting with statistician:

1. Define the type of data (continuous, ordinal, categorical, etc.)

A Few Examples of Statistical Tests

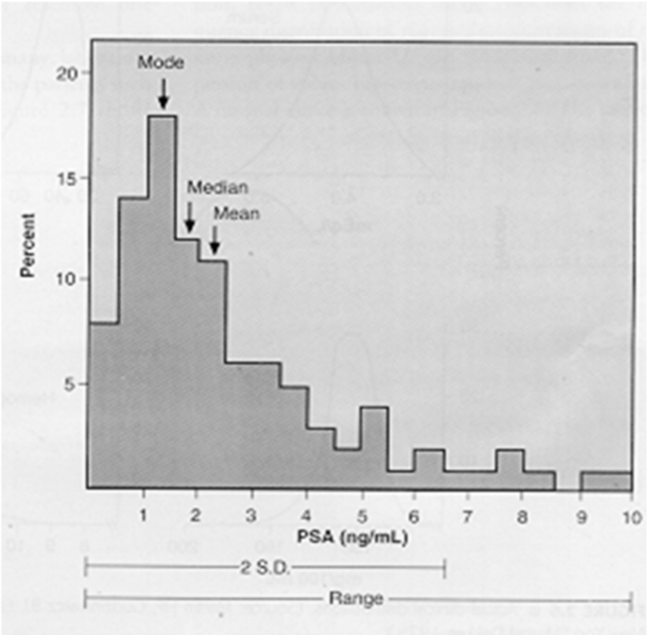
Test	Comparison	Principal Assumptions
Student's t test	Means of two groups	Continuous variable, normally distributed, equal variance
Wilcoxon rank sum	Medians of two groups	Continuous variable
Chi-square	Proportions	Categorical variable, more than 5 patients in any particular "cell"
Fisher's exact	Proportions	Categorical variable

Objectives - Concepts

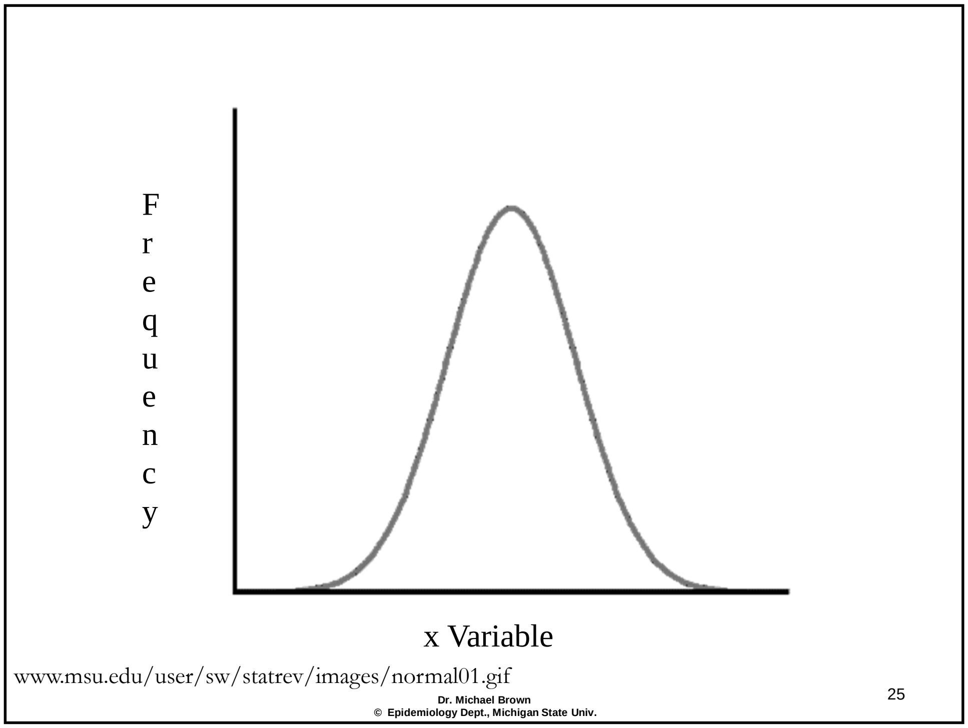
- Classification of data
- Distributions of variables
- Measures of central tendency and dispersion
- Criteria for abnormality
- Sampling
- Regression to the mean

Distributions of continuous variables

- A way to display the individual – to – individual variation in some clinical measure.
- Consider the example in Fletcher using PSA levels:



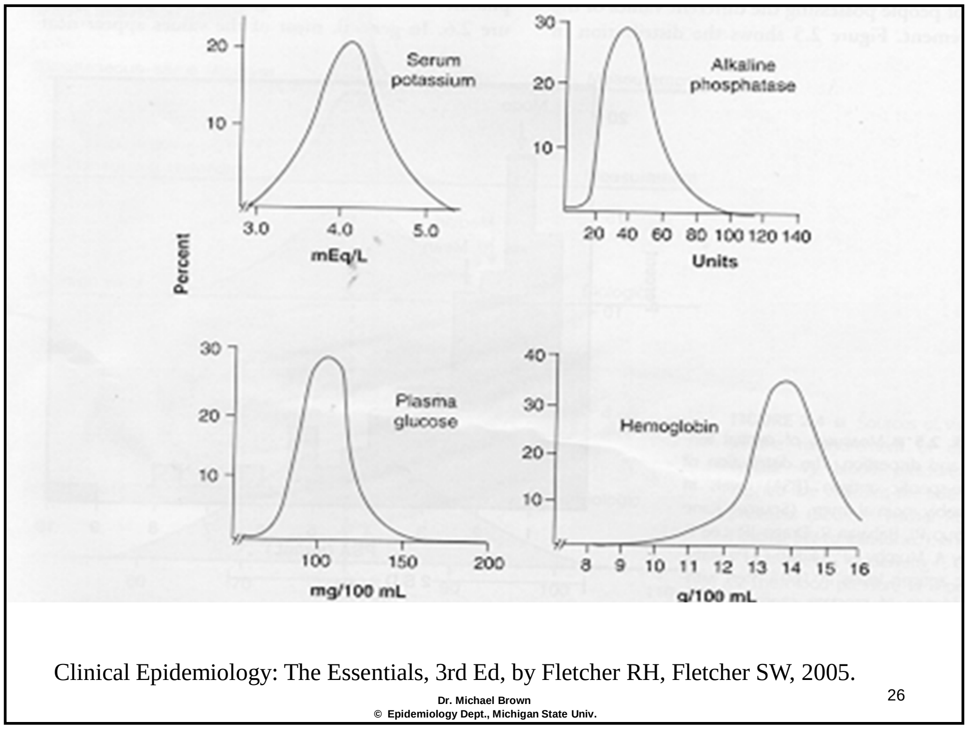
Clinical Epidemiology: The Essentials, 3rd Ed, by Fletcher RH, Fletcher SW, 2005.



www.msu.edu/user/sw/statrev/images/normal01.gif

Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.

25



Clinical Epidemiology: The Essentials, 3rd Ed, by Fletcher RH, Fletcher SW, 2005.

Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.

26

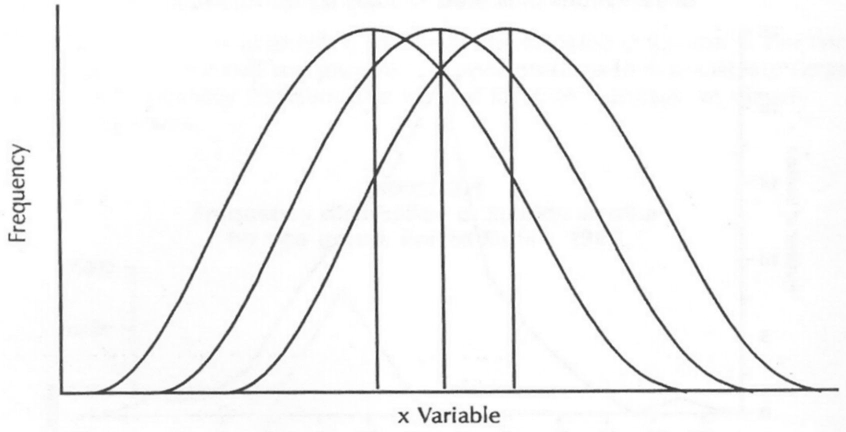
The “nicest” distribution

Is the normal, or Gaussian, distribution
– the “bell-shaped curve”.

If we want to summarize a frequency distribution, there are two major aspects to include:

- Central tendency
- Dispersion

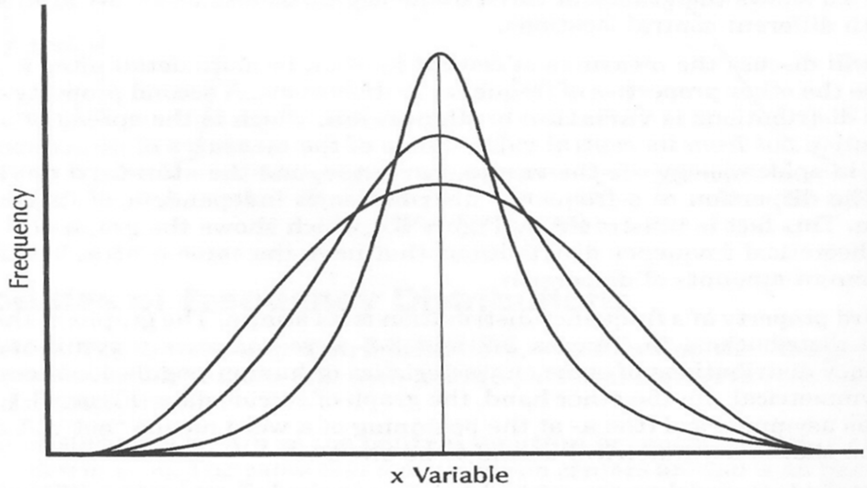
FIGURE 3.3
Three curves identical in shape with different central locations



Principles of Epidemiology, 2nd edition. CDC.

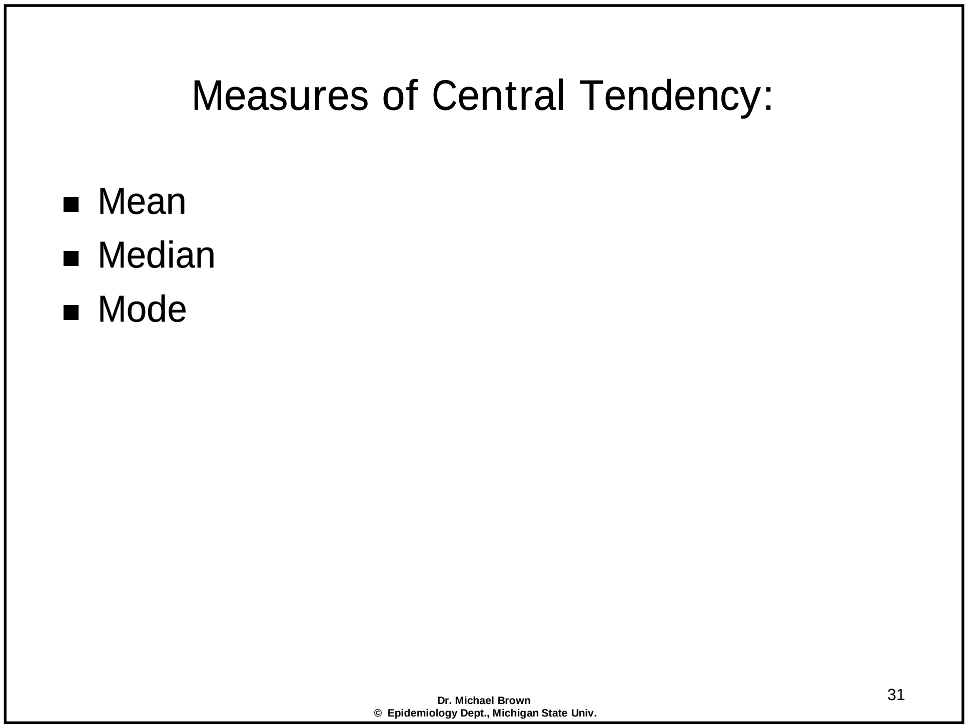
Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.

FIGURE 3.4
Three curves with same central location but different dispersion



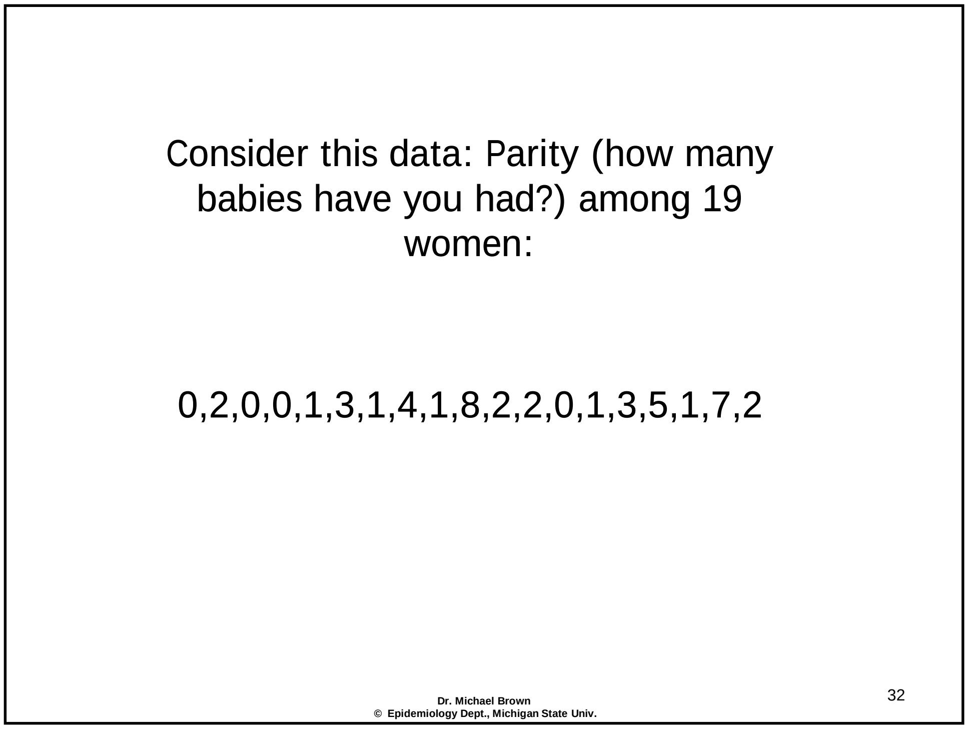
Principles of Epidemiology, 2nd edition. CDC.

Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.



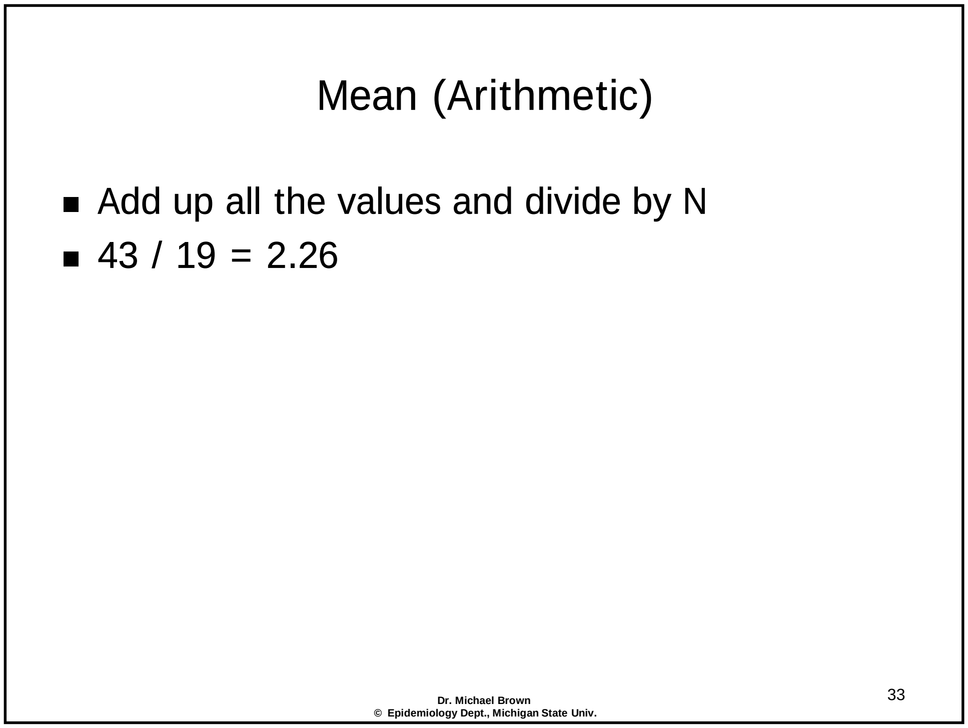
Measures of Central Tendency:

- Mean
- Median
- Mode



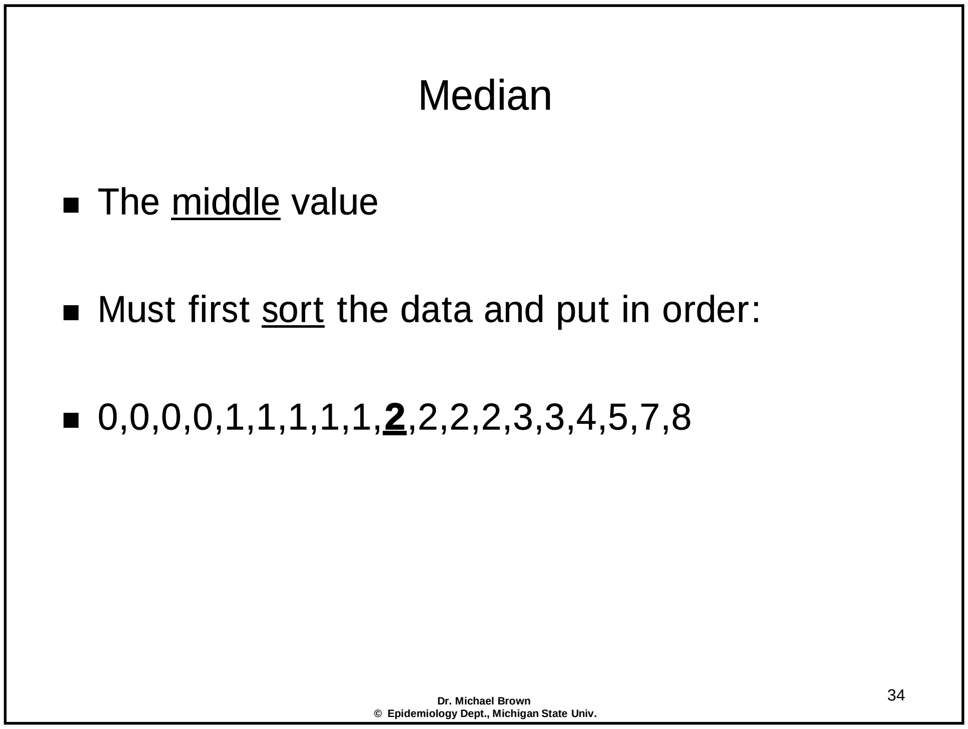
Consider this data: Parity (how many babies have you had?) among 19 women:

0,2,0,0,1,3,1,4,1,8,2,2,0,1,3,5,1,7,2



Mean (Arithmetic)

- Add up all the values and divide by N
- 43 / 19 = 2.26



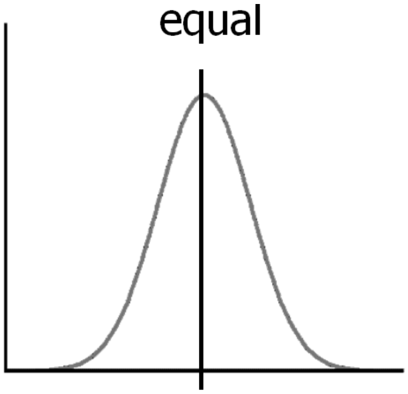
Median

- The middle value
- Must first sort the data and put in order:
- 0,0,0,0,1,1,1,1,1,**2**,2,2,2,3,3,4,5,7,8

Mode

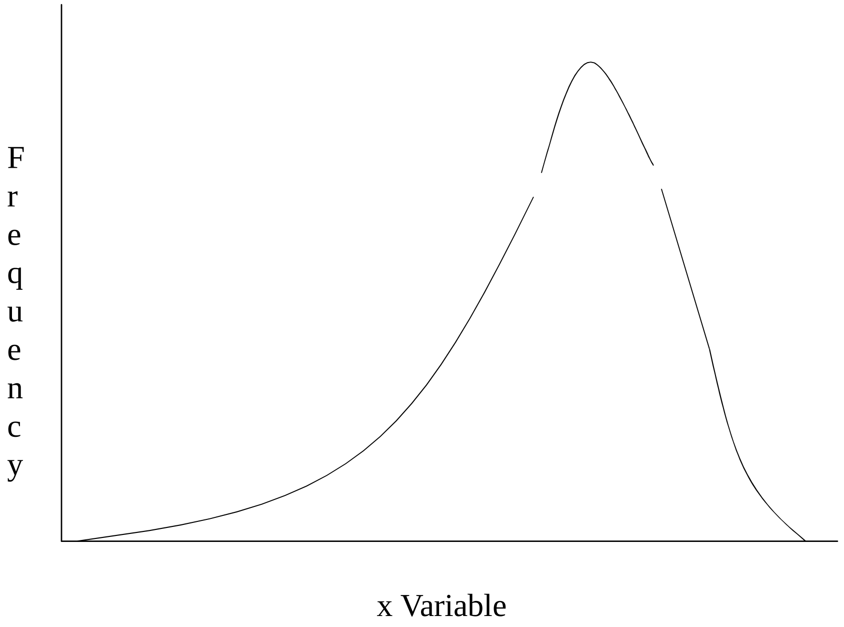
- The most common value
- 0,0,0,0,**1,1,1,1,1**,2,2,2,2,3,3,4,5,7,8

In a normal distribution, all three are equal



Parametric statistical methods assume a distribution with known shape

(i.e. normal or Gaussian distribution)



Quick Quiz Slide

- If the mode is “100” and the mean is “80” – what can you tell me about the median?

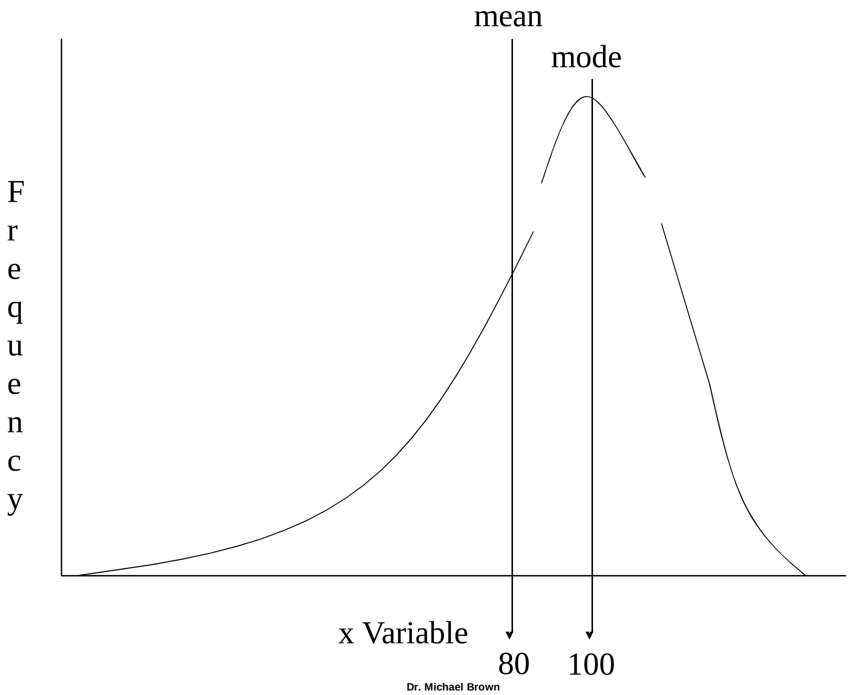
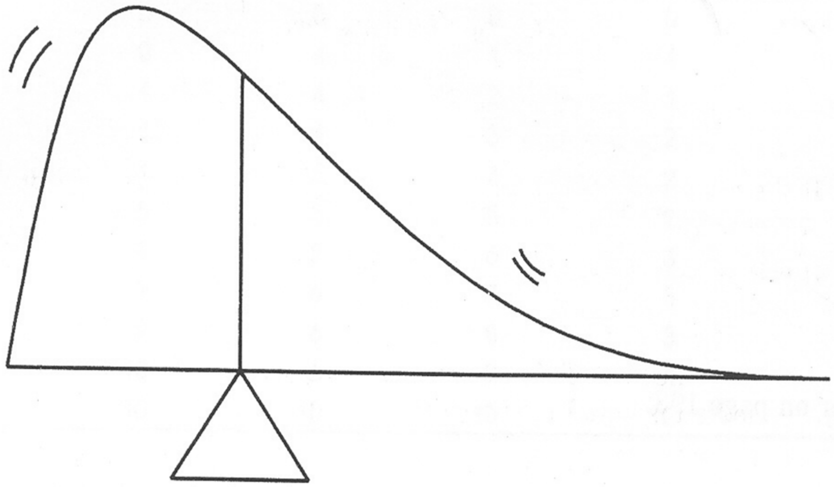


FIGURE 3.7
Mean is the center of gravity of the distribution



Dispersion

- Standard Deviation - most common measure used for normal or near normal distributions.
- Defined by a statistical formula, but remember that:
  - The mean +/- one SD contains about 2/3 of the observations.
  - the mean +/- 2 SD's includes about 95% of the observations.

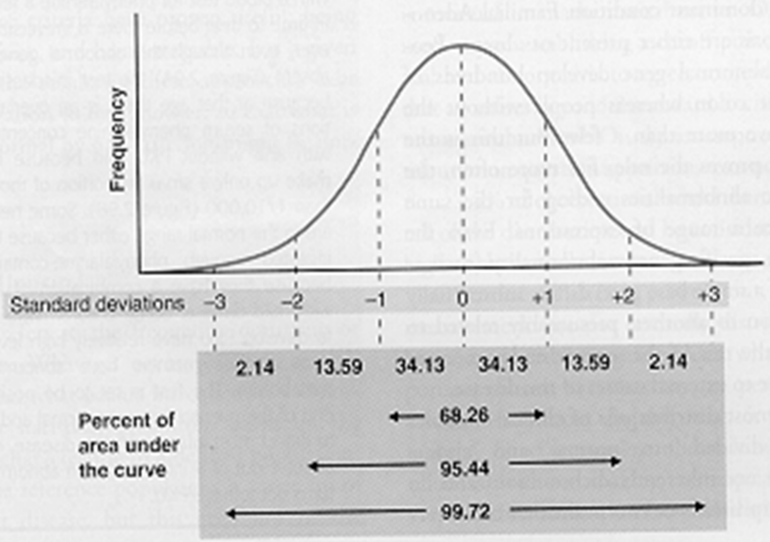
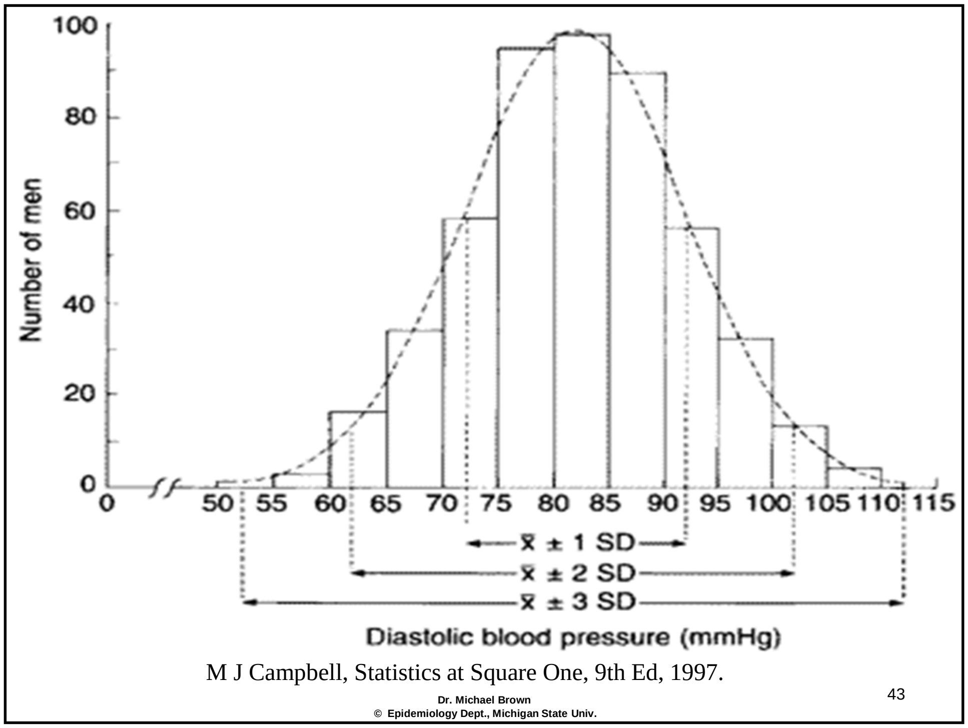
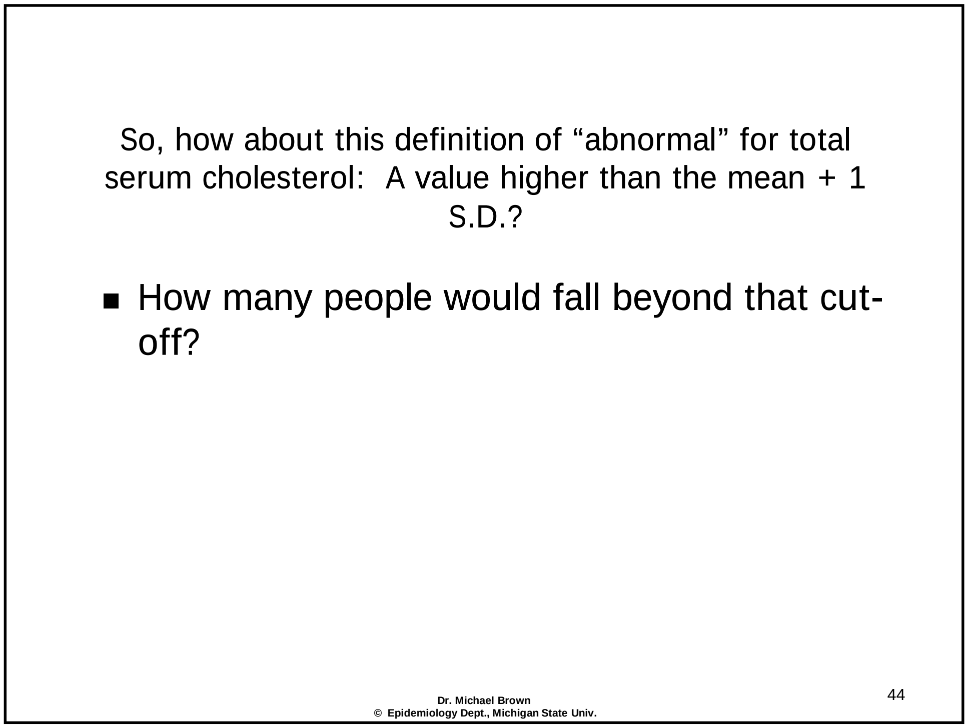


FIGURE 2.8 ■ The Normal (Gaussian) distribution.

Clinical Epidemiology: The Essentials, 3rd Ed, by Fletcher RH, Fletcher SW, 2005.



M J Campbell, Statistics at Square One, 9th Ed, 1997.



So, how about this definition of “abnormal” for total serum cholesterol: A value higher than the mean + 1 S.D.?

- How many people would fall beyond that cut-off?

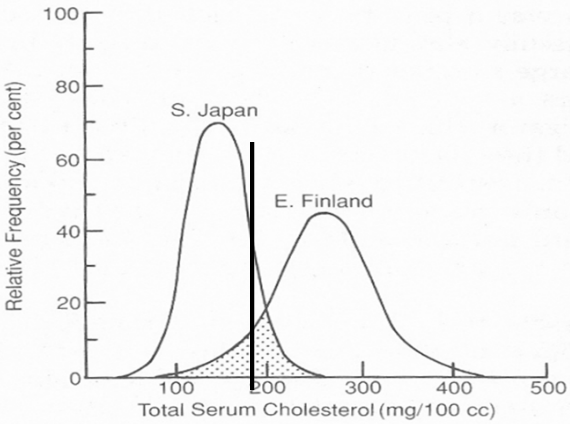
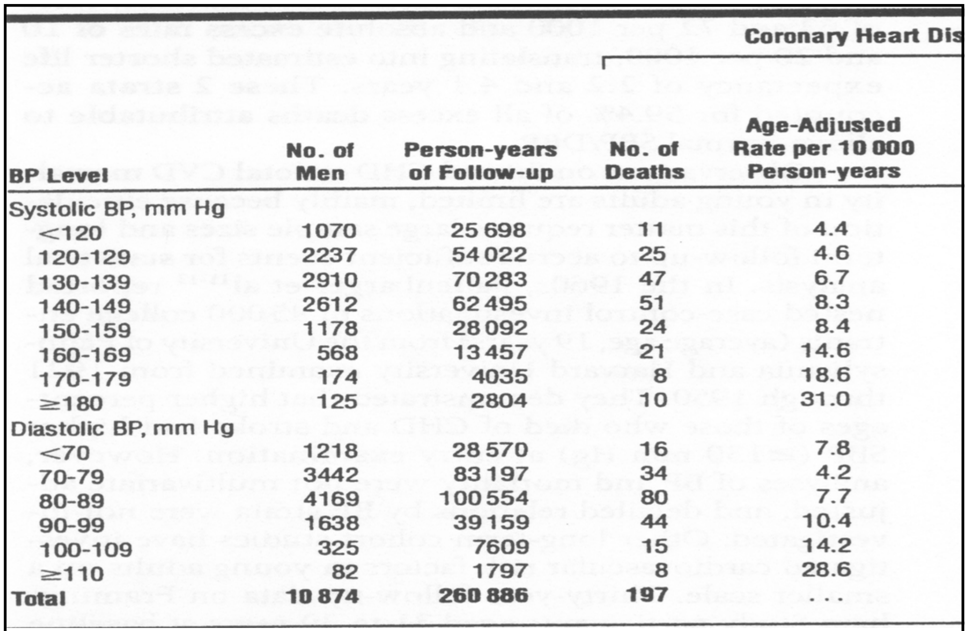


Fig. 5.1 The contrasting distributions of serum cholesterol in south Japan and eastern Finland.

Rose, G: The Strategy of Preventive Medicine; Oxford Press, 1998.

So what's the "best" definition of abnormality?

- Fletcher lists three:
  - Being unusual
    - Greater than 2 SD from mean
  - Sick
    - Observation regularly associated with disease
  - Treatable
    - Consider abnormal only if treatment of the condition represented by the measurement leads to improved outcome



Coronary Heart Dis

BP Level	No. of Men	Person-years of Follow-up	No. of Deaths	Age-Adjusted Rate per 10 000 Person-years
Systolic BP, mm Hg				
<120	1070	25 698	11	4.4
120-129	2237	54 022	25	4.6
130-139	2910	70 283	47	6.7
140-149	2612	62 495	51	8.3
150-159	1178	28 092	24	8.4
160-169	568	13 457	21	14.6
170-179	174	4035	8	18.6
≥180	125	2804	10	31.0
Diastolic BP, mm Hg				
<70	1218	28 570	16	7.8
70-79	3442	83 197	34	4.2
80-89	4169	100 554	80	7.7
90-99	1638	39 159	44	10.4
100-109	325	7609	15	14.2
≥110	82	1797	8	28.6
Total	**10 874**	**260 886**	**197**	. . .

Miura et al, Archives Int Med 2001; 161:1504.

If you were to design a study to define an abnormal DBP for adult females in the US, how would you do it?

- Measure DBP in every adult female in the US?
  - Then define abnormal as above 2 SD from mean?

Sampling

- Impossible to measure the BP of everyone, so must take measurements of a representative sample of subjects.
- Random sample
  - May miss important subgroup (ethnicity for example)
  - May need to obtain a larger sample from these important subgroups and select subjects at random within subgroup

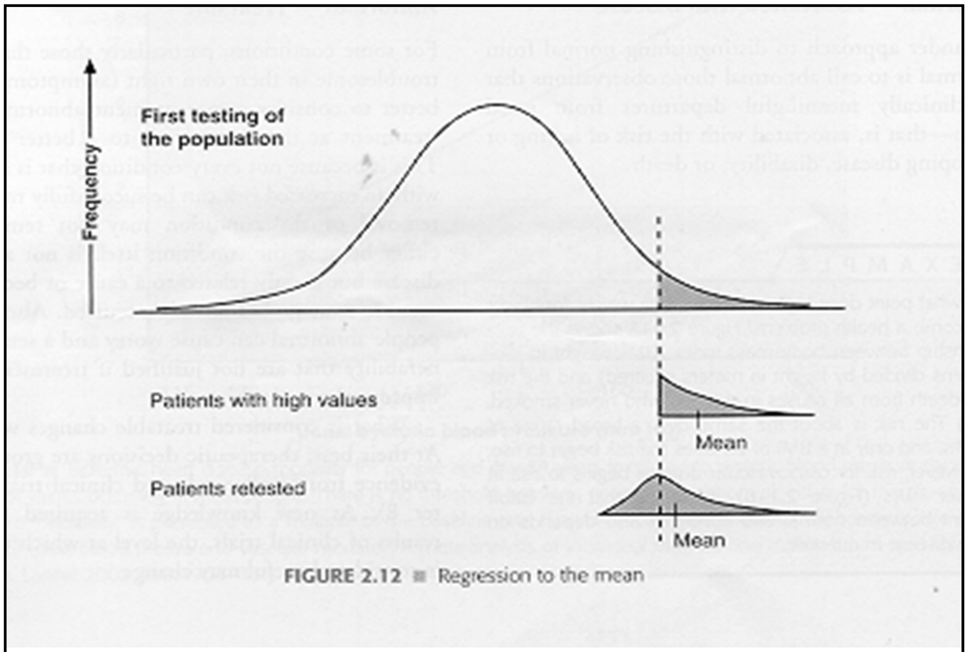
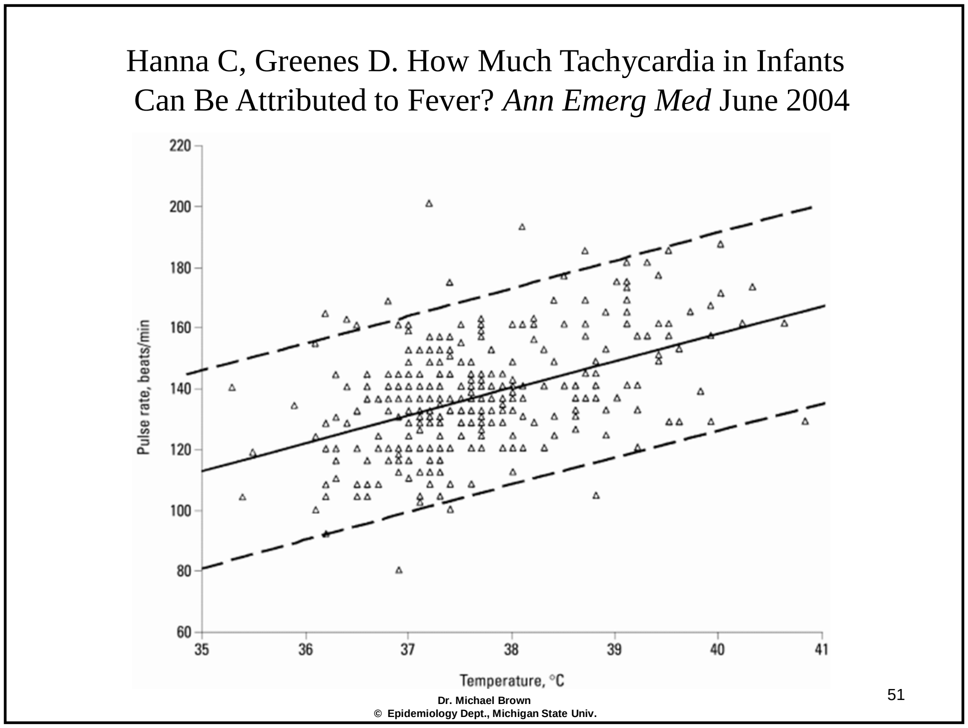


FIGURE 2.12 ■ Regression to the mean

Clinical Epidemiology: The Essentials, 3rd Ed, by Fletcher RH, Fletcher SW, 2005.



Hanna C, Greenes D. How Much Tachycardia in Infants Can Be Attributed to Fever? *Ann Emerg Med* June 2004

51

Dr. Michael Brown
© Epidemiology Dept., Michigan State Univ.

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes - Descriptive Statistics

Normality, Abnormality, and Medical Measurement

Mat Reeves, PhD

Objectives

I. Understand the use of different criteria to define normality and abnormality

II. Understand the rationale for sampling
   a. Random vs. systematic error
   b. Statistical inference – estimation and hypothesis testing

III. Understand the origins of variation in clinical data
   a. Biological and measurement variation
   b. Distinguish between inter- and intra- person/observer variability

IV. Understand the difference between validity and reliability
   a. Definition of validity and reliability
   b. Measures of validity (sensitivity, specificity)
   c. Measures of reliability (kappa, intra-class correlation)
   d. Describe combinations of validity and precision (target diagrams)

V. Statistical aspects of variability
   a. Measures of variation (Standard deviation (SD), variance)
   b. Measures of agreement (Correlation, Kappa) – understand the logic behind these 2 measures and how to interpret them.

VI. Statistical aspects of clinical data
   a. Classification of data (categorical [nominal, ordinal], interval [discrete, continuous]
   b. Distributions (normal, left and right skew)
   c. Measures of central tendency (mean, median, mode)
   d. Measures of dispersion (SD, range, inter-quartile range IQR)
   e. Regression to the mean

I. Normality vs. Abnormality

Chapter 2 in FF is informative and easy to follow. The online lecture summarizes some of the key concepts.

II. Sampling

It is very difficult if not impossible to obtain data from every member of a population – case in point is the U.S. census which attempts to contact every household in the country. In year 2000, its response rate was only 65%. So a more practical approach is to take a *representative sample* of the underlying population and then draw inferences about the population from this sample. Taking a sample always involves an element of random variation or error – *sampling statistics* are essentially about characterizing the nature and magnitude of this random error.

Random error can be defined as *the variation that is due to "chance"* and is an inherent feature of "sampling" and statistical inference. However, in clinical medicine we also recognize that random error can occur due to the process of measurement and/or the biological phenomenon itself. For example, a given blood pressure measurement may vary because of random error in the measurement tool (sphygmomanometer) and due to the "natural" biological variation in the underlying blood pressure.

While the field of statistics is essentially concerned with the characterization of random error, the degree to which any data is affected by *systematic error or bias* is probably more important. Systematic error can be defined as *any process that acts to distort data or findings from their true value*. In epidemiology, we classify bias as either *selection bias*, *measurement bias* or *confounding bias* (and we will refer to these sources of bias frequently during this course and in EPI-547). It is important to note that traditional "statistics" for the most part addresses only random error and NOT systematic bias. Thus, a significant *P* value or a precise confidence interval cannot tell you whether the underlying data is accurate i.e., unbiased.

Statistical Inference is the process whereby one draws conclusions regarding a population from the results observed in a sample taken from that population. There are two different but complementary categories of statistical inference: estimation and hypothesis testing. *Estimation* is concerned with estimating the specific value of an unknown population parameter, while *hypothesis testing* is concerned with making a decision about a hypothesized value of an unknown population parameter. These concepts will be further explored in Lecture 4 and in Chapter 10 of FF.

III. Variation in Clinical Data

Biological and Measurement Variation

Clinical information is no different from any other source of data - it has inherent variability which can create substantial difficulties. All types of clinical data whether concerning the patient's history, physical exam, laboratory results, or response to treatment may change even over the shortest of time intervals. In the broadest sense, variation may be grouped into two categories: variation in the actual entity being measured (*biological variation*) and variation due to the measurement process (*measurement variation*).

1. *Biologic Variation:*
The causes and origins of biologic variation are endless; variation derives from the dynamic nature of physiology, homeostasis and pathophysiology, as well as genetic differences and differences in the way individuals react to changing environments such as those induced by disease and treatment. Biologic variation can be further sub-divided into within (*intra-person*)

and between *(inter-person*) variability. For example, your blood pressure shows a high degree of intra-person variability- it is changing by the hour or even minute in response to many stimuli, such as time of day, posture, physical exertion, emotions, and that last shot of expresso. Biologic variation also occurs because of differences between subjects (inter-person variability). Fortunately there is enough variation between individuals, compared to the degree of within-person variability, that after several repeat blood pressure measurements it is possible to determine the typical (average) blood pressure value of an individual patient and classify them as to their hypertension status i.e., normal, pre-hypertension, or hypertension (Stage I, II or III). Different variables have different amounts of within and between subject variation which can have important clinical consequences.

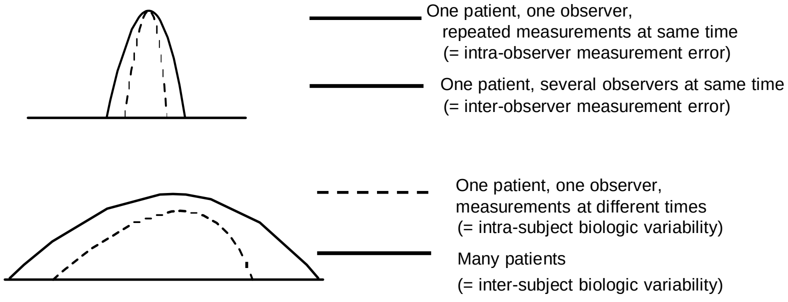
Regardless of the source of biological variation, its net effect is to add to the level of random error in any measurement process. A common method of reducing the impact of biological variation is to take repeated measurements of a variable or phenomenon - as in the above example of blood pressure. Finally, note that the presence of biological variation is *sine qua non* for epidemiologists to define factors that are associated with disease or outcomes i.e., risk factors. If everyone in the population has the same value or outcome then it is impossible to study the disease process.

2. *Measurement Variation*:
Measurement variation is derived from the measurement process itself. It may be caused by inaccuracy of the instrument (*instrument error*) or the person operating the test (*operator error*). Measurement variation can introduce both *random* error into the data as well as *systematic* error or bias, especially when the test requires some human judgment. Systematic differences between laboratories is one reason that it is vital for each lab to establish its own "reference" ranges. Variation due to different observers reading the same test (e.g., radiologists reading the same x-ray) is referred to as *inter-observer variability*, whereas variability resulting from the same observer reading the same test (e.g., one radiologist reading the same x-ray at different times) is referred to as *intra-observer variability*. Approaches to reduce the impact of measurement bias include the use of specific operational standards e.g., assay standards for laboratory instruments, or the use of explicit operational procedures as in the example of blood pressure measurement (i.e., seated position, appropriate cuff size, identification of specific Korotkoff sounds, repeated measurements etc). Random variation due to measurement variation can again be lessened by taking repeated measurements of a variable or phenomenon.

Note that all variation is additive, so that the net observed variation is a result of the culmination of various individual sources. This is shown nicely by the following figure adapted from the Fletcher text:

Figure 2.1. Sources of variation in measuring diastolic blood pressure

Cumulative Sources of Variation –
Measurement of DBP (from Fletcher)

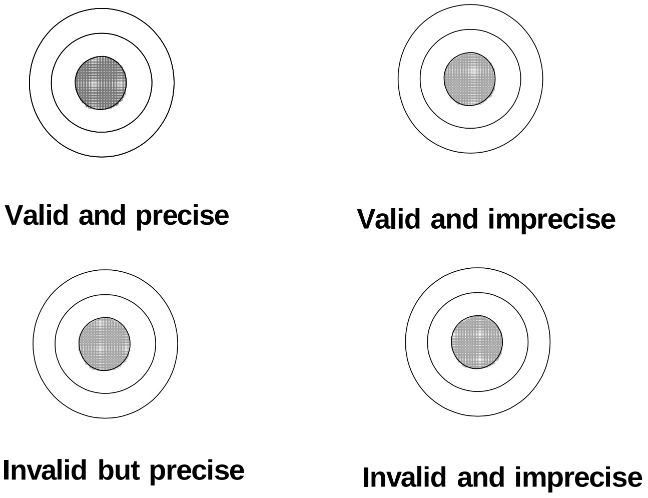

One patient, one observer, repeated measurements at same time (= intra-observer measurement error)
One patient, several observers at same time (= inter-observer measurement error)
One patient, one observer, measurements at different times (= intra-subject biologic variability)
Many patients (= inter-subject biologic variability)

IV. Validity and Reliability

Validity refers to the degree to which a measurement process tends to measure what is intended i.e., how accurate is the measurement. A valid instrument will, on average, be close to the underlying true value, and is therefore free of any *systematic error* or *bias*. Graphically, validity can be depicted as a series of related measurements that cluster around the true value (see Fig 2.2). For some clinical data, such as laboratory variables, validity can easily be determined by comparing the observed measurement to an accepted "*gold standard*". However, for other clinical data, such as pain, nausea or anxiety there are no obvious gold standard measures available. In this case, it is common to develop instruments that are thought to measure some specific phenomena or *construct*. These constructs are then used to develop a *clinical scale* that can be used to measure the phenomenon in practice. The validity of the instrument or scale can then be evaluated in terms of *content validity* (i.e., the extent to which the instrument includes all of the dimensions of the construct being measured – this is also called *face* validity), *construct validity* (i.e., the degree to which the scale correlates with other known measures of the phenomenon) and *criterion validity* (i.e., the degree to which the scale predicts a directly observable phenomenon).

For dichotomous data, validity is usually expressed in terms of *sensitivity* and *specificity* (see lecture 5). There are several different statistical methods for expressing the validity of continuous data, including presenting the mean and standard deviation of the difference between the surrogate measure and the gold standard, as well as correlation and regression analysis.

Fig 2.2 <u>Schematic representation of validity and reliability (for you to enjoy completing…)</u>



Reliability (or *reproducibility*) refers to the extent that repeated measurements of a phenomenon tend to yield the same results - regardless of whether they are correct or not. There is therefore no comparison to a reference or gold standard measure. Reliability refers to the *lack of random error* - the degree of reliability or is inversely related to the amount of random error - the more error the less precise the instrument. Graphically, reliability can be depicted as the degree to which a series of related measurements cluster together (Fig 2.2).

Random variation can be classified according to whether there is one or multiple observers or instruments i.e., *intra-observer variability* vs. *inter-observer variability*, respectivel*y*. Using the target analogy of Fig 2.2, inter-observer reliability refers to the scatter from different observers shooting at the same target, while intra-observer reliability refers to the scatter of shots from one shooter.

For measurements that do not involve a direct observation e.g., self-administered questionnaires, reliability can be assessed using the *test-retest method*, where respondents answer the same question at two different times. This approach measures a form of intra-observer reliability, where the respondent is acting as the same observer on two separate occasions. The exact statistical approach used to quantify reliability depends on the type of data measured – Kappa for categorical data and intra-class correlation for interval data.

V. Statistical aspects of variability

A. Measures of variation

Data variability can be quantified by various standard statistical measures of dispersion such as variance, standard deviation (SD), and range.

1. *Variance and Standard Deviation*

The variability or precision of a measurement is expressed by the standard deviation (SD). The SD represents the absolute value of the average difference of individual values from the mean, and is calculated by taking the square root of the variance.

$$SD = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Assuming a normal distribution, one standard deviation either side of the mean includes 68% of the total number of observations, while 2 SD's include 95%.

> Example:
>
> The SD of serum cholesterol is 15 mg/dl. Americans now have an average cholesterol value of 205 mg/dl, thus the values for the middle 68% of the population would be expected to vary from 190 to 220 mg/dl (mean ± 1 SD)

B. Measures of agreement

1. *Correlation (r)*

The reliability of a continuous measurement (i.e., interval data) can be expressed by the correlation coefficient (*r*) between two sets of measurements. The correlation coefficient *(r)* measures the strength of the linear relationship between two continuous variables: *r* ranges from -1 to +1, with zero representing no relationship.

If information can be obtained on the actual true values, then the correlation can be regarded as a test of *validity* between the "truth" and an imperfect measurement. However, true values are rarely available, so in most cases, the correlation between two measures assesses *reliability* i.e., the extent to which the results can be replicated by two measurements.

If measures are obtained from two observers, then the extent of agreement between the two would reflect the between-observer variability (or reliability). However, it should be noted that it is possible to have high values of r, yet have little direct agreement between two observers or instruments. For example, in measuring blood cholesterol, a perfect r (1.0) can occur if laboratory A results are always exactly 10 mg/dl points higher than laboratory B.

The correlation co-efficient is also commonly used as a measure of reliability in *test-retest studies* - where the same instrument is applied to the same population at a later time (a measure of intra-rater or within-person variability).

2. Categorical data - Kappa

For categorical or qualitative data, reliability can be characterized using the *kappa statistic (k).* Kappa has the useful property of correcting for the degree of chance in the overall level of agreement, and is therefore preferred over other measures like the commonly used *overall percent agreement*. The ability of kappa to adjust for chance agreement is especially important in clinical data, because the prevalence of the particular condition being evaluated affects the likelihood that observers will agree purely due to chance. This chance agreement must be adjusted for, otherwise false reassurances can occur. As an example of this phenomenon, if 2 people each repeatedly toss a coin, there are only 4 possible results i.e., HH (i.e., head, head), TT (i.e., tail, tail), HT, and TH. The probability (p) of each result is ¼, so the overall agreement between the two coins (due to chance alone) is 0.5 (i.e., sum of p(HH) and p(TT)). The influence of the underlying prevalence of the attribute or condition being measured on the overall percent agreement is shown in the following table:

Overall percent agreement due to chance for a binary attribute

Prevalence of the attribute	**Overall percent agreement***
0.1	82%
0.3	58%
0.5	50%
0.7	58%
0.9	82%

* P(e) calculated by multiplying the marginal totals of a 2x2 table

The kappa (k) statistic is calculated as:

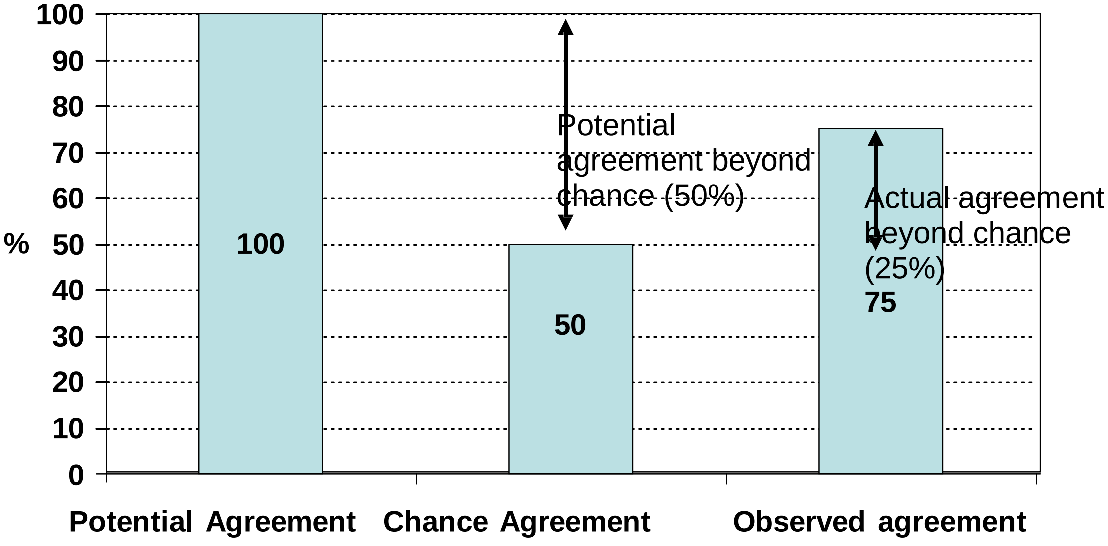
$$k = \frac{P_o - P_e}{1 - P_e}$$

where, P_o = the total proportion of observations on which there is agreement.
P_e = the proportion of agreement expected by chance alone.

Thus k is the ratio of the actual agreement attributable to the reproducibility of the observations (i.e., $P_o - P_e$), compared to the maximum possible value ($1 - P_e$). Or,

$$k = \frac{\text{Actual agreement beyond chance}}{\text{Potential agreement beyond chance}}$$

The following diagram explains the logic of the kappa statistic. In this example chance corrected agreement (Kappa) = (0.75-0.50)/(1-0.50) = 0.25/0.50 = 50%.



The simplest clinical application of Kappa is in the measurement of inter-rater agreement whereby two observers evaluate the same series of patients and classify them according to some particular dichotomous condition (e.g., disease present or absent). As an example, the following data is generated from 2 radiologists who independently reviewed 150 mammograms and classified each patient as to whether they had an abnormality:

OBSERVER B	OBSERVER A		
	Yes	No	TOTALS
Yes	69	15	84
No	18	48	66
TOTALS	87	63	150

Observer A thought 87 patients (or 58%) had an abnormality, while Observer B thought 84 patients did (i.e., 56%), but they agreed only 78% of the time (the observed proportion of agreement (P_o) is calculated as (69 + 48)/150 = 0.78 or 78%). However, the proportion of agreement expected due to chance (P_e) was 0.51 or 51%, which is estimated by calculating the “expected” numbers in cells a and d from the product of the marginal totals (i.e., 87 x 84/150 = 48.72 [for cell a], and 63 x 66/150 = 27.72 [for cell d], and then calculating the proportion of agreement by dividing the sum of these two cells by the total number of subjects i.e., 48.72 + 27.72 / 150, which equals 0.5096 or 51%). Thus k can be estimated as:

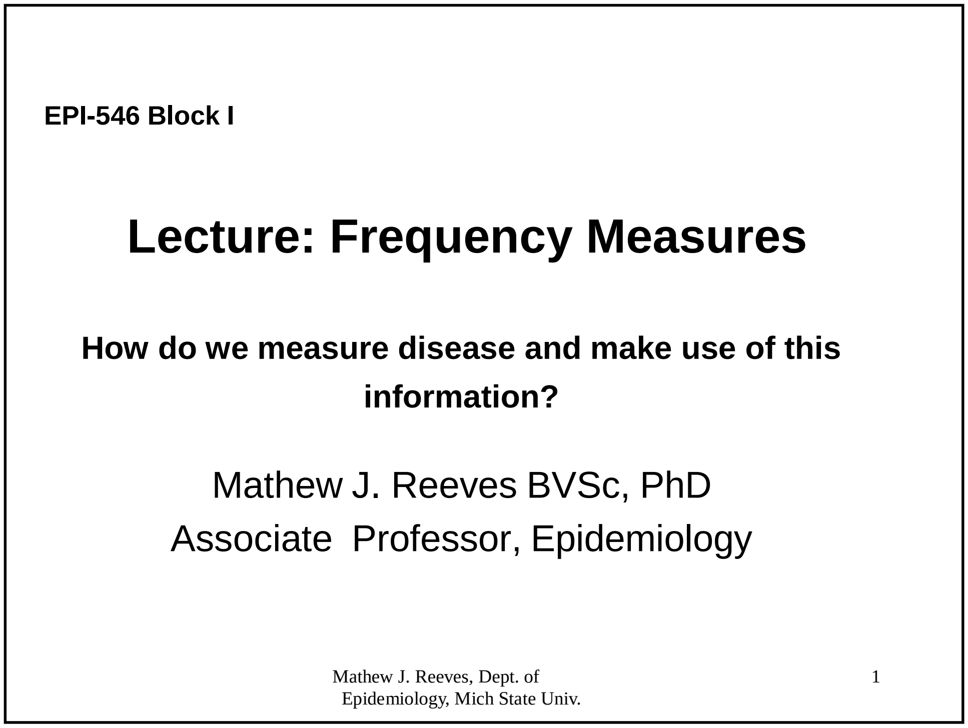
$$k = \frac{P_o - P_e}{1 - P_e} = \frac{0.78 - 0.51}{1 - 0.51} = \frac{0.27}{0.49} = 0.55 \text{ or } 55\%$$

Like the correlation coefficient, kappa varies in value from -1 to +1, however the interpretation is different. A value of zero denotes agreement that is no better than chance, while a negative value denotes agreement that is worse than chance (i.e., fulminant disagreement!). The following guide to interpreting the strength of agreement shown by kappa has been developed. In our example - 55% represents a moderate degree of agreement between the two radiologists.

Value of k	**Strength of agreement**
<0	Poor
0 - 0.20	Slight
0.21 - 0.40	Fair
0.41 - 0.60	Moderate
0.61 - 0.80	Substantial
0.81 - 1.0	Almost perfect

VI. Statistical aspects of clinical data

See Chapter 2 in FF and online lecture.

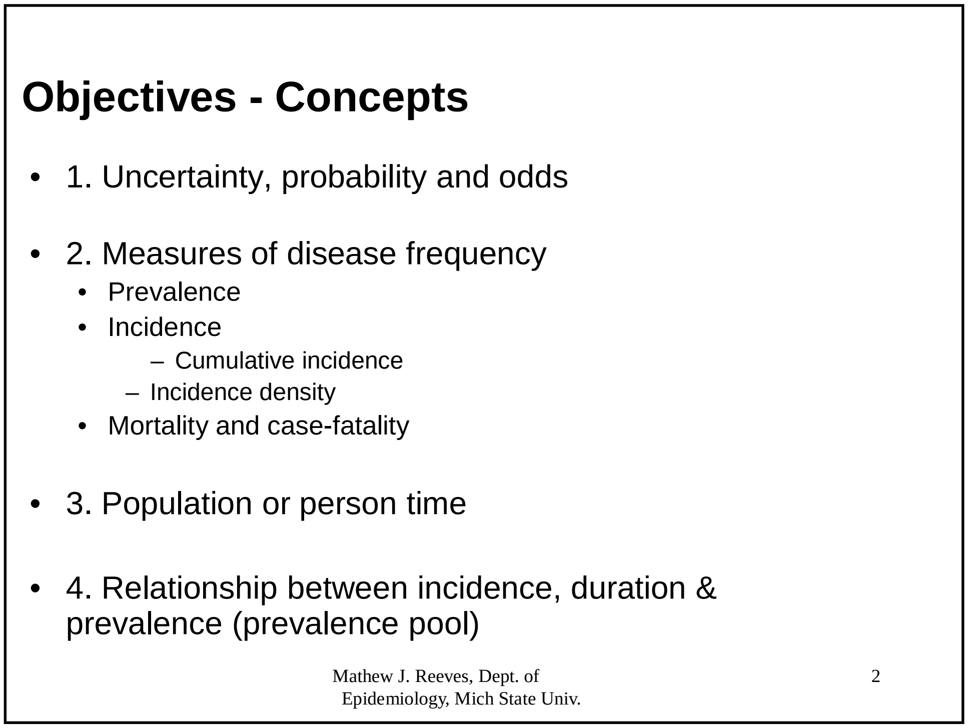


EPI-546 Block I

Lecture: Frequency Measures

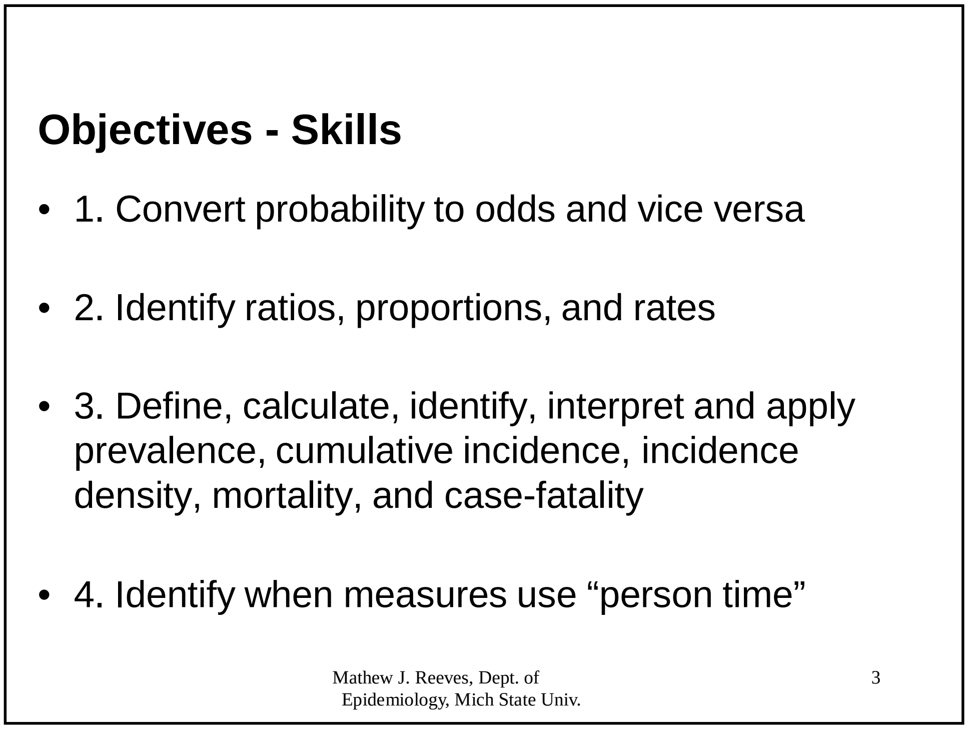
How do we measure disease and make use of this information?

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology



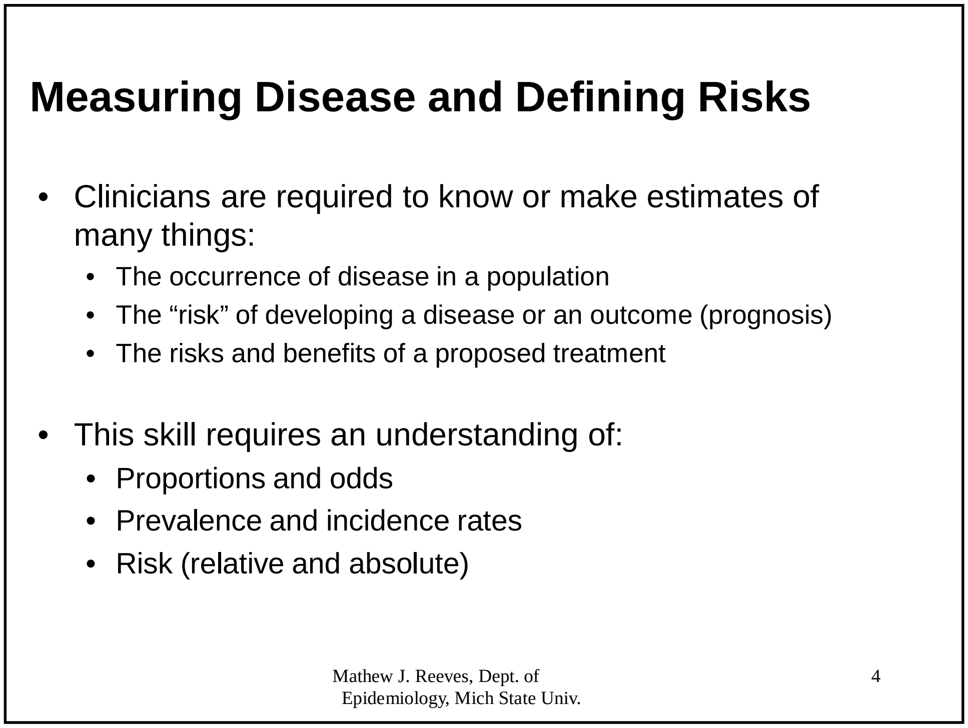
Objectives - Concepts

- 1. Uncertainty, probability and odds
- 2. Measures of disease frequency
  - Prevalence
  - Incidence
    - Cumulative incidence
    - Incidence density
  - Mortality and case-fatality
- 3. Population or person time
- 4. Relationship between incidence, duration & prevalence (prevalence pool)



Objectives - Skills

- 1. Convert probability to odds and vice versa
- 2. Identify ratios, proportions, and rates
- 3. Define, calculate, identify, interpret and apply prevalence, cumulative incidence, incidence density, mortality, and case-fatality
- 4. Identify when measures use "person time"



Measuring Disease and Defining Risks

- Clinicians are required to know or make estimates of many things:
  - The occurrence of disease in a population
  - The "risk" of developing a disease or an outcome (prognosis)
  - The risks and benefits of a proposed treatment
- This skill requires an understanding of:
  - Proportions and odds
  - Prevalence and incidence rates
  - Risk (relative and absolute)

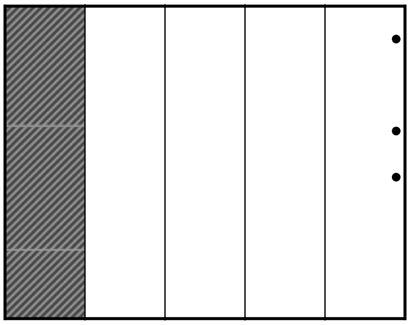
Uncertainty

- Medicine isn’t an exact science, uncertainty is ever present
- Uncertainty can be expressed either:
  - Qualitatively using terms like ‘probable’, ‘possible’, ‘unlikely’
    - Study: Docs asked to assign prob. to commonly used words:
      - *‘Consistent with’* ranged from 0.18-0.98
      - *‘Unlikely’* ranged from 0.01 to 0.93
  - Quantitatively using probabilities (P)
    - Advantage: explicit interpretation, exactness
    - Disadvantage: may force one to be more exact that is justified!

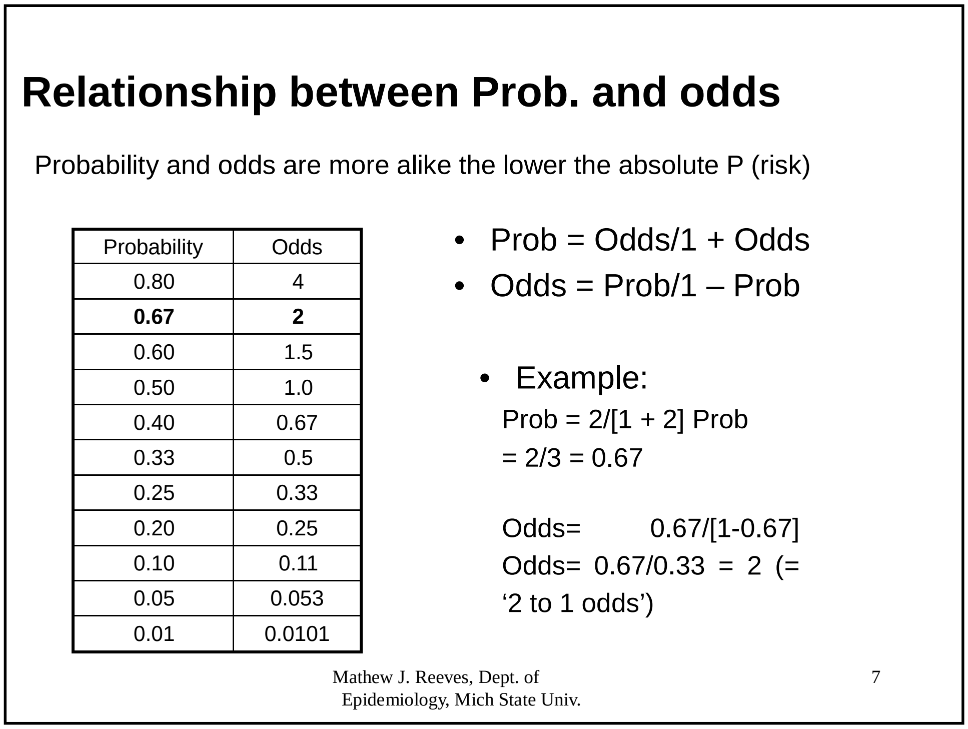
Probability vs. Odds

Probability (P) or “risk” of having an event
Odds = ratio of the probability of having an event to the probability of not having the event or P / (1 – P)

Example: 1 out of 5 patients suffer a stroke…….



- P = 1/5 = 0.2 or 20%
- Odds = (P) / (1-P)
- Odds = 0.2 / 0.8 or 1:4 or “one to four”



Relationship between Prob. and odds

Probability and odds are more alike the lower the absolute P (risk)

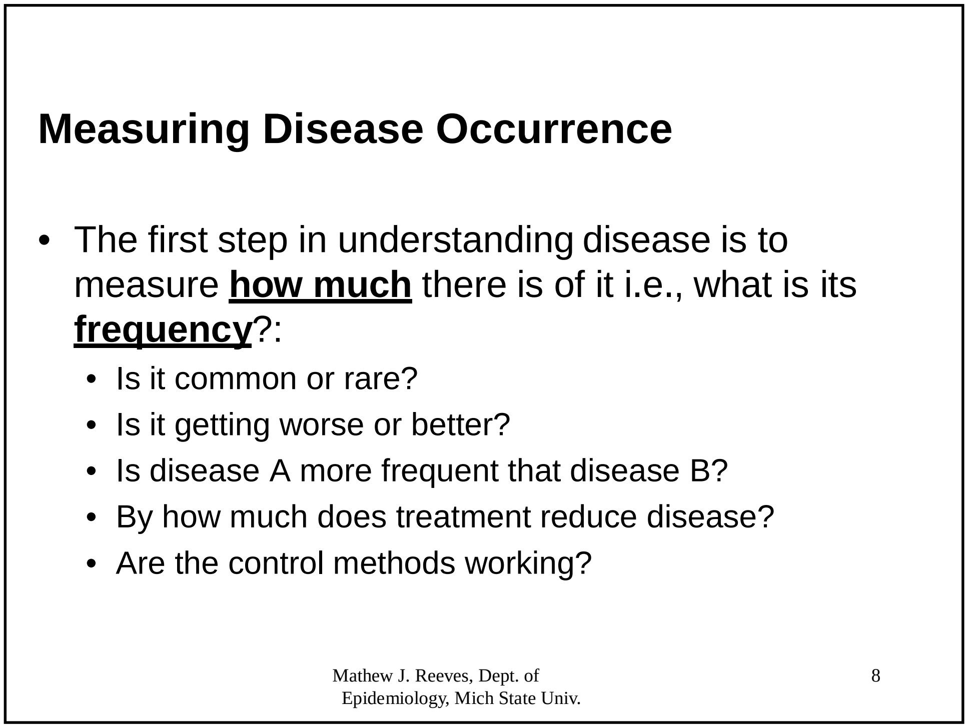
Probability	Odds
0.80	4
0.67	**2**
0.60	1.5
0.50	1.0
0.40	0.67
0.33	0.5
0.25	0.33
0.20	0.25
0.10	0.11
0.05	0.053
0.01	0.0101

- Prob = Odds/1 + Odds
- Odds = Prob/1 – Prob

- Example:

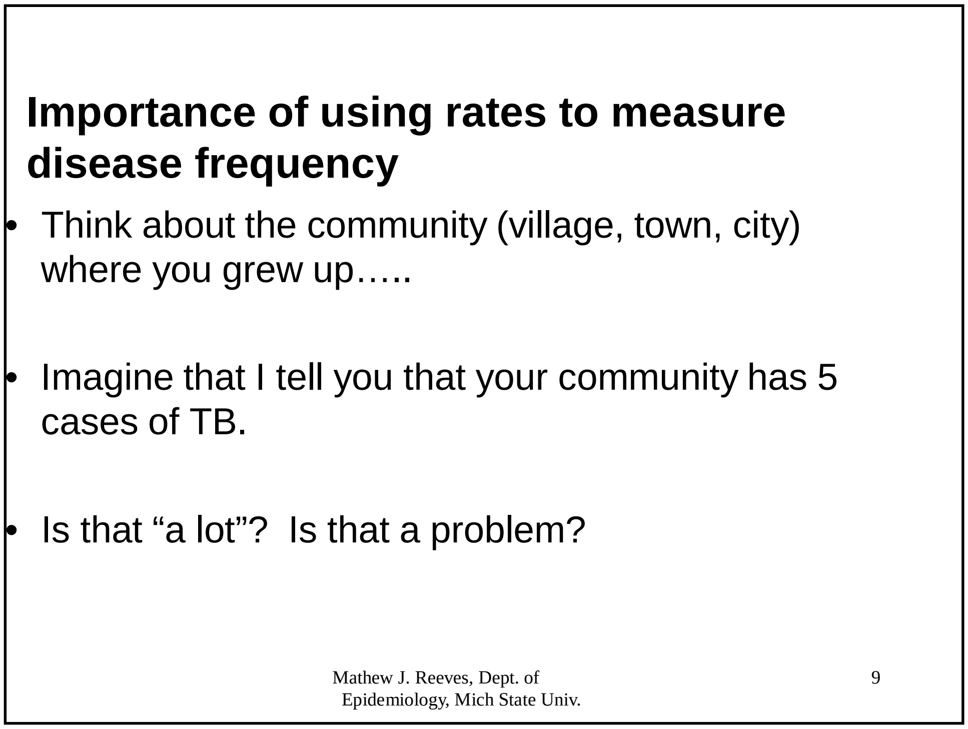
Prob = 2/[1 + 2] Prob = 2/3 = 0.67

Odds= 0.67/[1-0.67]
Odds= 0.67/0.33 = 2 (= ‘2 to 1 odds’)



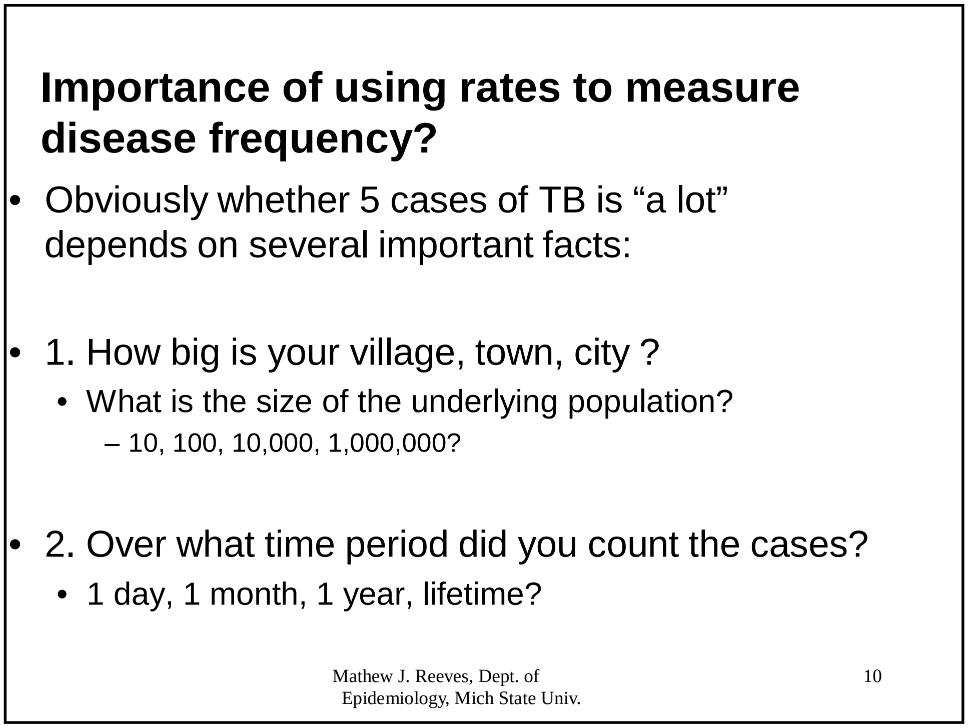
Measuring Disease Occurrence

- The first step in understanding disease is to measure **how much** there is of it i.e., what is its **frequency**?:
  - Is it common or rare?
  - Is it getting worse or better?
  - Is disease A more frequent that disease B?
  - By how much does treatment reduce disease?
  - Are the control methods working?



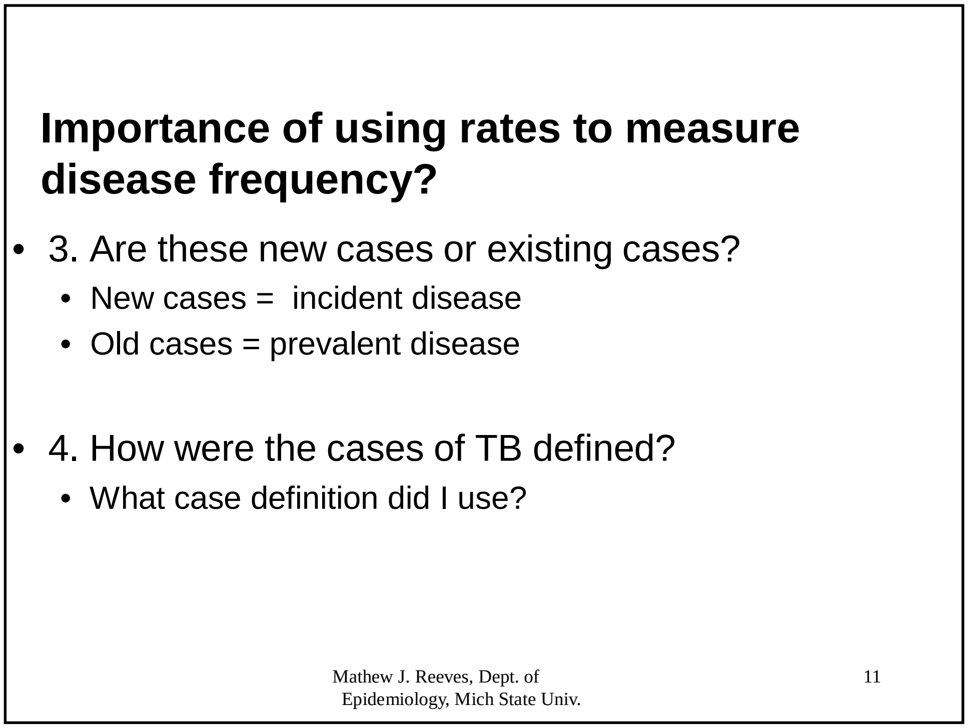
Importance of using rates to measure disease frequency

- Think about the community (village, town, city) where you grew up…..
- Imagine that I tell you that your community has 5 cases of TB.
- Is that “a lot”? Is that a problem?



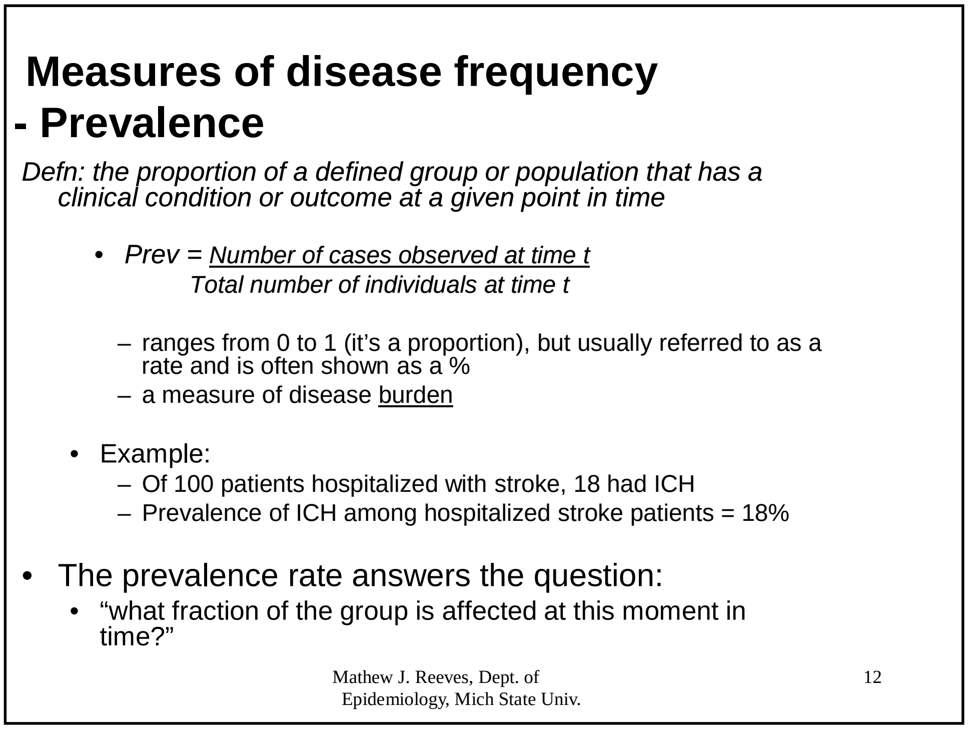
Importance of using rates to measure disease frequency?

- Obviously whether 5 cases of TB is “a lot” depends on several important facts:
- 1. How big is your village, town, city ?
  - What is the size of the underlying population?
    - – 10, 100, 10,000, 1,000,000?
- 2. Over what time period did you count the cases?
  - 1 day, 1 month, 1 year, lifetime?



Importance of using rates to measure disease frequency?

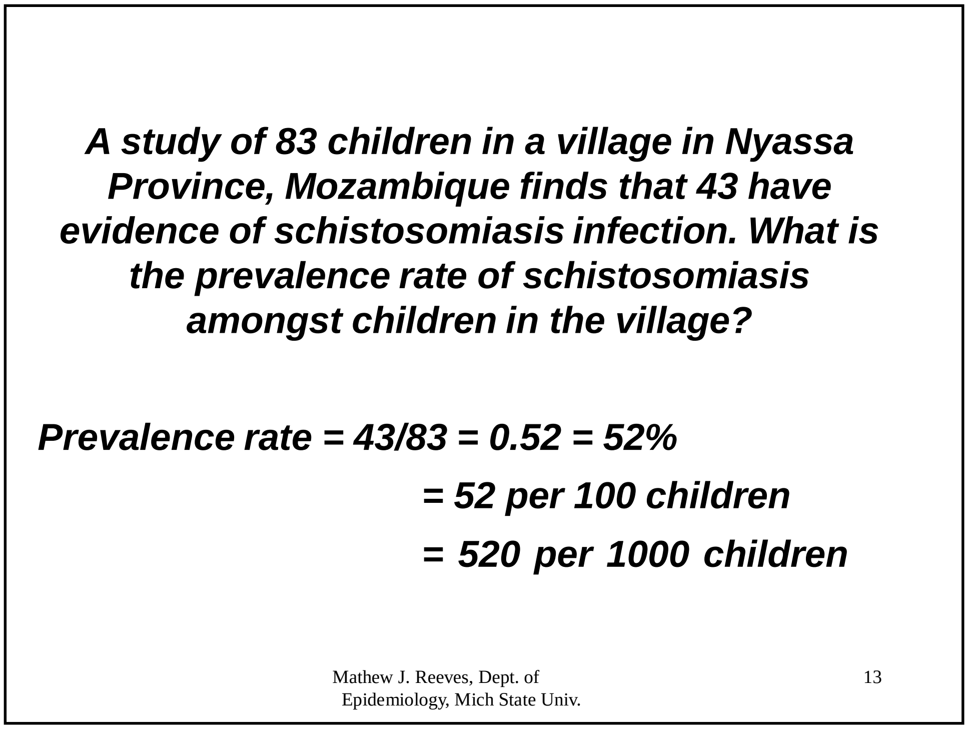
- 3. Are these new cases or existing cases?
  - New cases = incident disease
  - Old cases = prevalent disease
- 4. How were the cases of TB defined?
  - What case definition did I use?



Measures of disease frequency - Prevalence

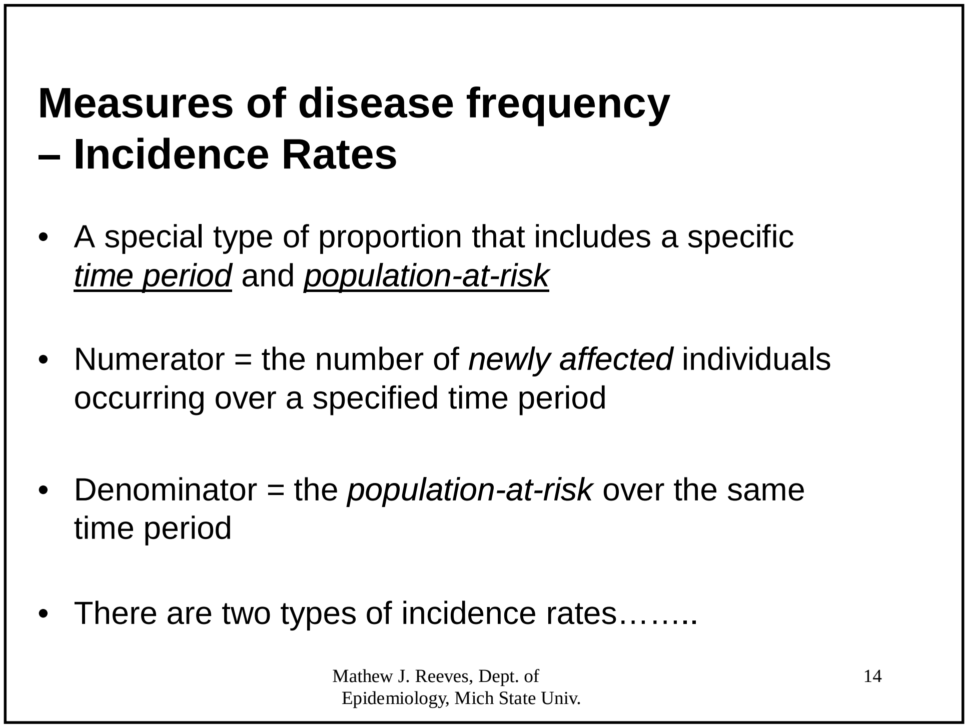
Defn: the proportion of a defined group or population that has a clinical condition or outcome at a given point in time

- $\textit{Prev} = \dfrac{\underline{\textit{Number of cases observed at time t}}}{\textit{Total number of individuals at time t}}$
  - ranges from 0 to 1 (it’s a proportion), but usually referred to as a rate and is often shown as a %
  - a measure of disease burden
- Example:
  - Of 100 patients hospitalized with stroke, 18 had ICH
  - Prevalence of ICH among hospitalized stroke patients = 18%
- The prevalence rate answers the question:
  - “what fraction of the group is affected at this moment in time?”



A study of 83 children in a village in Nyassa Province, Mozambique finds that 43 have evidence of schistosomiasis infection. What is the prevalence rate of schistosomiasis amongst children in the village?

Prevalence rate = 43/83 = 0.52 = 52%
= 52 per 100 children
= 520 per 1000 children



Measures of disease frequency – Incidence Rates

- A special type of proportion that includes a specific *time period* and *population-at-risk*

- Numerator = the number of *newly affected* individuals occurring over a specified time period

- Denominator = the *population-at-risk* over the same time period

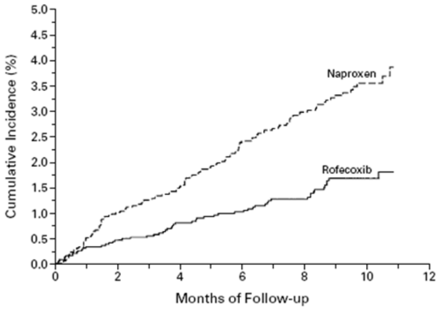
- There are two types of incidence rates……..

Cumulative Incidence Rate (CIR)

Defn: the proportion of a defined at-risk group or population that <u>develops</u> a new clinical condition or outcome <u>over</u> a given time period.

- *CIR=* $\dfrac{\textit{Num. of newly disease indv. for a specific time period}}{\textit{Total number of population-at-risk for same time period}}$
- Measures the proportion of at-risk individuals who develop a condition or outcome over a specified time period
- Ranges from 0 to 1 (so it's a proportion!) but called a rate because it includes time period and population-at-risk
- Must be accompanied by a specified time period to be interpretable - because the CIR must increase with time
  - 7-day CIR of stroke following TIA = 5%
  - 90-day CIR of stroke following TIA = 10%

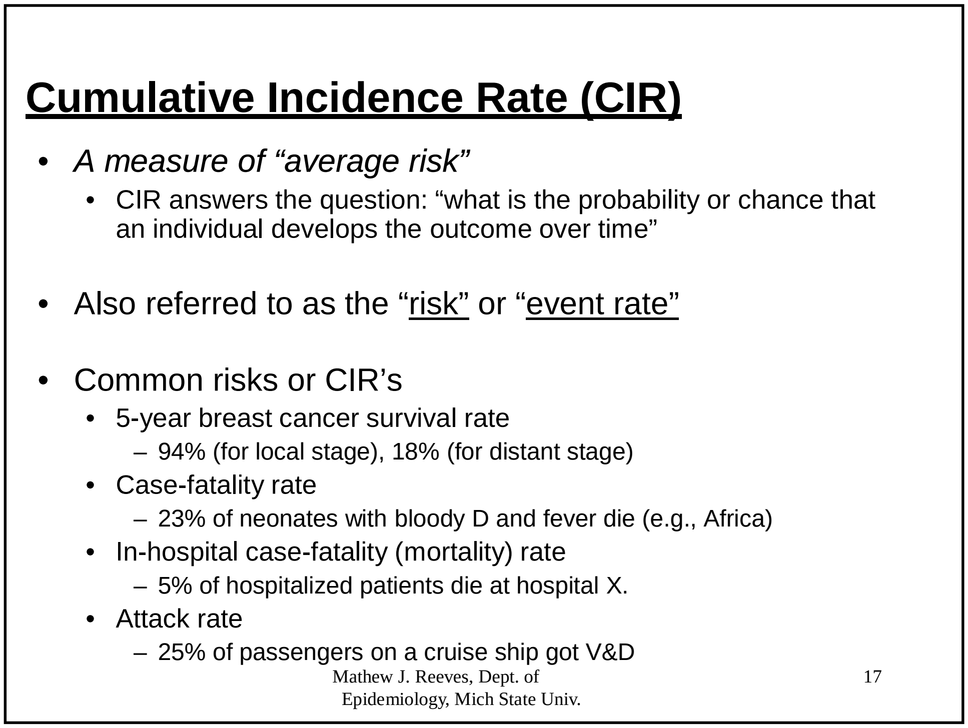
Cumulative Incidence of GI side effects for Rofecoxib (VIOXX) vs. Naproxen - The VIGOR Trial (Bombardier NEJM 2000)



No. at Risk

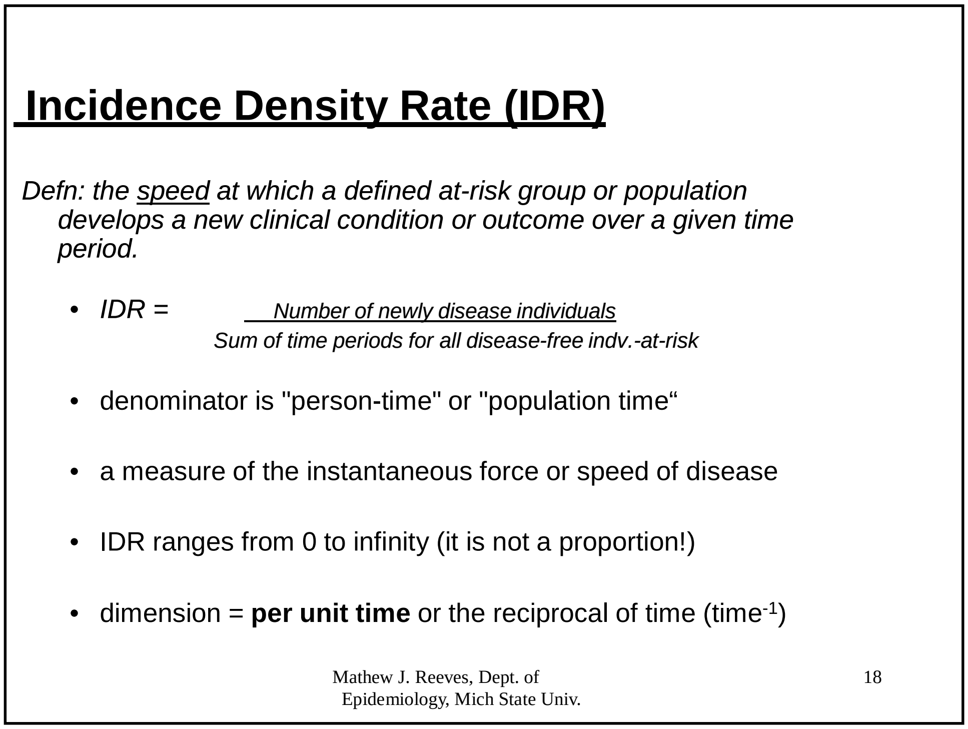
Rofecoxib	4047	3641	3402	3180	2806	1073	533
Naproxen	4029	3644	3389	3163	2796	1071	513

Figure 1. Cumulative Incidence of the Primary End Point of a Confirmed Upper Gastrointestinal Event among All Randomized Patients.



Cumulative Incidence Rate (CIR)

- *A measure of "average risk"*
  - CIR answers the question: "what is the probability or chance that an individual develops the outcome over time"
- Also referred to as the "risk" or "event rate"
- Common risks or CIR's
  - 5-year breast cancer survival rate
    - 94% (for local stage), 18% (for distant stage)
  - Case-fatality rate
    - 23% of neonates with bloody D and fever die (e.g., Africa)
  - In-hospital case-fatality (mortality) rate
    - 5% of hospitalized patients die at hospital X.
  - Attack rate
    - 25% of passengers on a cruise ship got V&D



Incidence Density Rate (IDR)

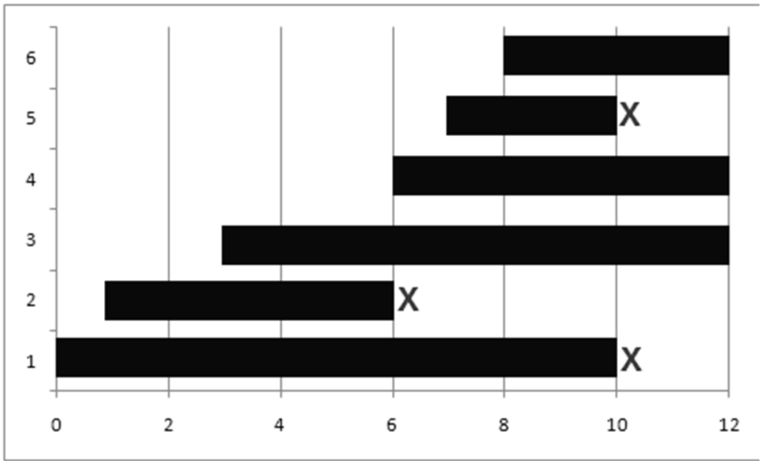
Defn: the speed at which a defined at-risk group or population develops a new clinical condition or outcome over a given time period.

- $IDR = \dfrac{\textit{Number of newly disease individuals}}{\textit{Sum of time periods for all disease-free indv.-at-risk}}$
- denominator is "person-time" or "population time“
- a measure of the instantaneous force or speed of disease
- IDR ranges from 0 to infinity (it is not a proportion!)
- dimension = **per unit time** or the reciprocal of time (time^{-1})

The Concept of “person-time”

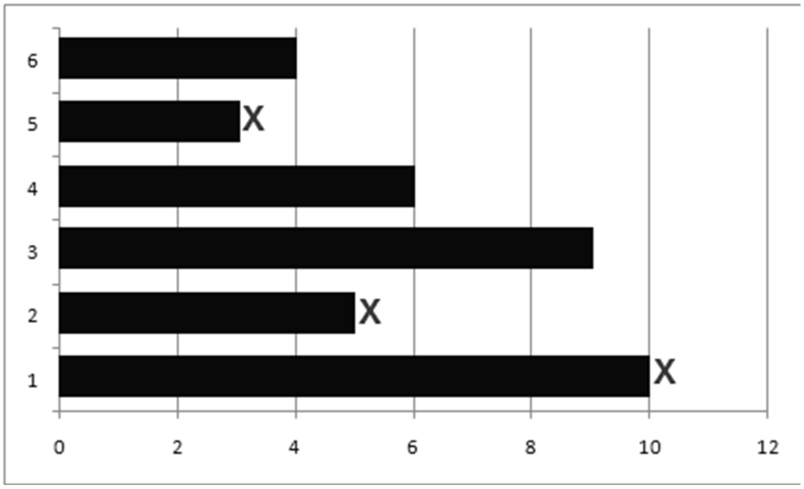
- the sum of the disease-free time experience for individuals at risk in the population
- Concept: 100 people followed for 6 months have same person-time experience as 50 people followed for a year.
  - 100 x 0.5 = 50 person-years
  - 50 x 1.0 = 50 person-years
- How do you calculate?
  - Simply add up the disease-free time experiance
  - 100 subjects followed for 12 months = 100 person-years
  - If one new case developed after 6 months then person time = 99.5 person- years
- Person time can be measured with whatever scale that makes the most sense i.e., person-days, person-weeks, person-months, person-years (PY)

Enrollment of 6 subjects in a 12-month study (X = event)



Question: What is the incidence rate?

What is the incidence rate?



Incidence Density Rate (IDR)

- *A measure of the "speed" that disease is occurring*
  - IDR answers the question: "At what rate are new cases of disease occurring in the population"
- Common IDR's
  - Mortality rate (Vital Statistics)
    - Lung CA mortality rate = 50 per 100,000 PY
    - Breast CA mortality rate = 15 per 100,000 PY
  - Disease Incidence Rates
    - IDR of neonatal diarrhea = 280 per 1,000 child weeks
  - Disease specific IDR rates
    - Calculated for specific sub-sets defined by age, gender or race
      - Black Men: Lung CA incidence rate = 122 per 100,000 PY
      - Wh. Female: Lung CA incidence rate = 43 per 100,000 PY

Approximating "person-time"

- Unless the population is small or the number of events rare, in most cases it is usually sufficient to approximate the person time

- Lung Cancer Incidence and Mortality. Michigan Cancer Registry 2005:

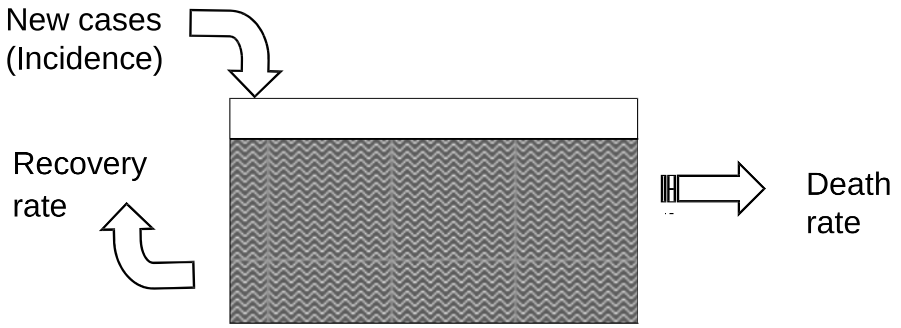
	Population	Cases	Deaths	Incidence/ 100,000 PY's	Mortality/ 100,000 PY's
Total	10,125,000	7,681	5,789	75.8	57.1
Male	4,975,000	4,218	3,179	84.8	63.9
Female	5,150,000	3,463	2,610	67.2	50.7

Population size estimated on July 1st

Choice Between CIR and IDR?

	CIR	IDR
Population	Closed	Open
Starting Point	Fixed	Open
Type of outcome	Single (death)	Multiple (URT infection)
Fluctuation in rates	Stable (cancer stats)	Highly variable (outbreaks)
Example	Fixed cohort (medical school class)	Open cohort (RCT)

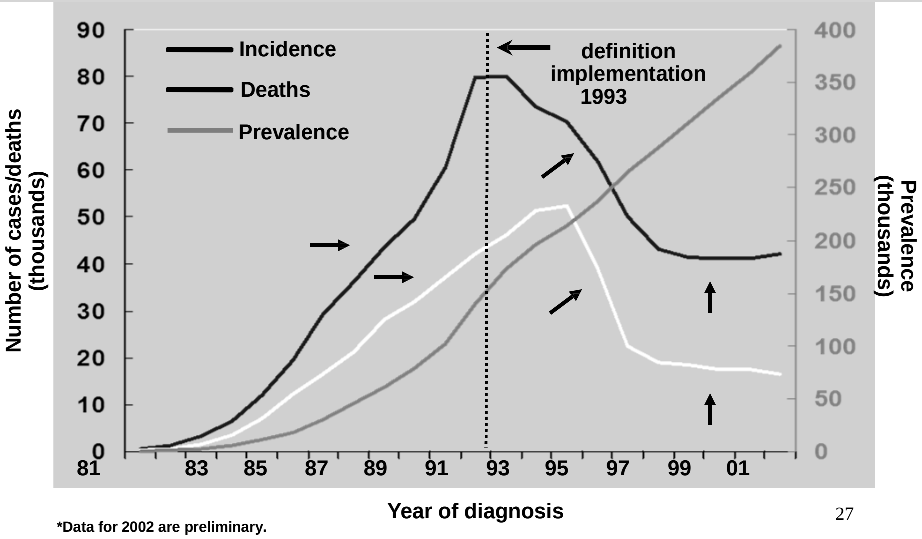
Concept of the Prevalence “Pool”



Relationship between Prevalence and Incidence

- Prevalence is a function of:
  - the incidence of the condition, and
  - the average duration of the condition
    - duration is influenced in turn by the recovery rate and mortality rate
- Prev ~ Incidence x Duration
- This relationship explains why….
  - Arthritis is common (“prevalent”) in the elderly
  - Rabies is rare.
  - Influenza is only common during epidemics.

Trends in AIDS Incidence, Prevalence, and Deaths, 1981 – 2002*



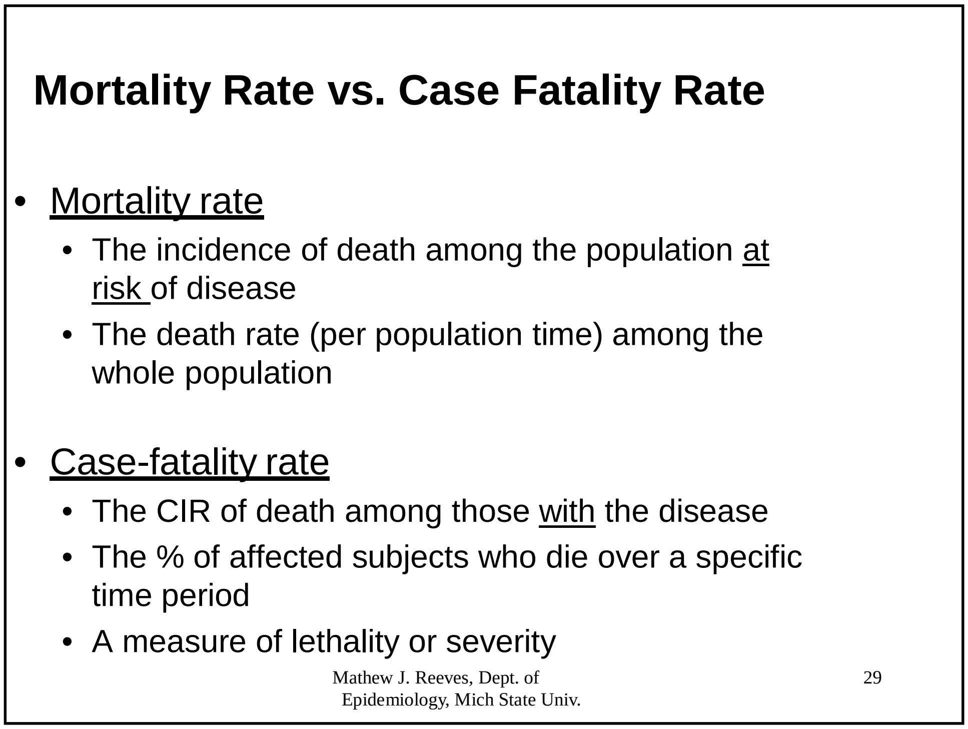
*Data for 2002 are preliminary.

SOURCE: HIV/AIDS Reporting System, CDC/NCHSTP.

Mortality (=death) rates

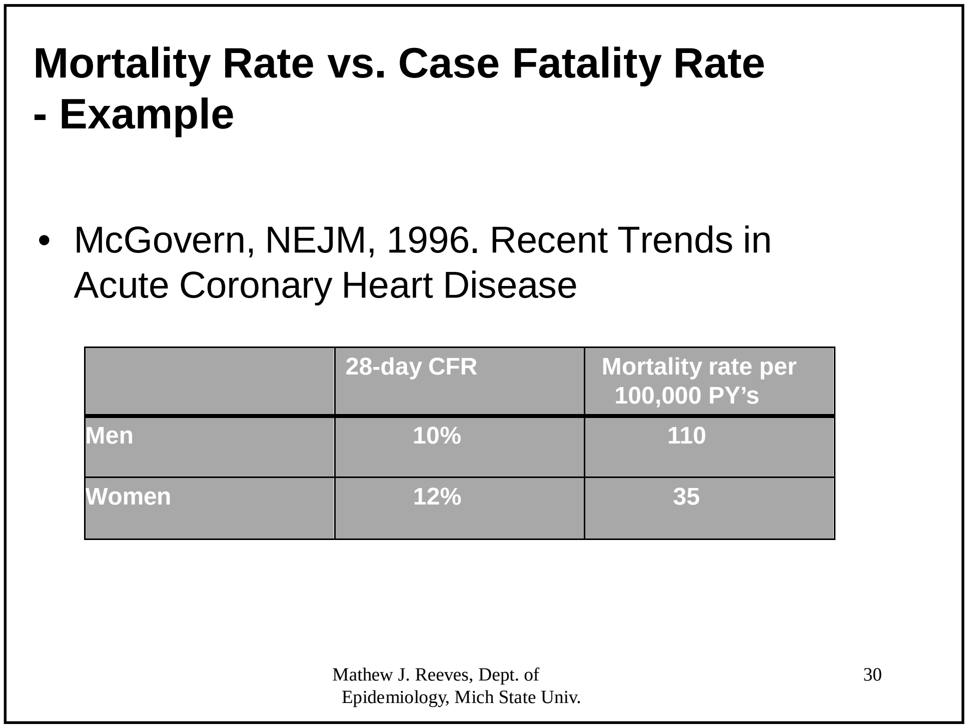
- The frequency of death in a defined population during a specified time period

- Mortality rate (all causes)=
  - *CIR*= $\frac{\textit{Number of deaths during a specific time period}}{\textit{Total number of population-at-risk for same time period}}$
  - *IDR* = $\frac{\textit{Number of deaths during a specific time period}}{\textit{Sum of time periods for all individuals-at-risk}}$



Mortality Rate vs. Case Fatality Rate

- Mortality rate
  - The incidence of death among the population at risk of disease
  - The death rate (per population time) among the whole population
- Case-fatality rate
  - The CIR of death among those with the disease
  - The % of affected subjects who die over a specific time period
  - A measure of lethality or severity



Mortality Rate vs. Case Fatality Rate - Example

- McGovern, NEJM, 1996. Recent Trends in Acute Coronary Heart Disease

	28-day CFR	Mortality rate per 100,000 PY's
Men	10%	110
Women	12%	35

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes - Frequency Measures

Mat Reeves BVSc, PhD

Objectives:

I. Understand the concept of uncertainty and quantify using probability and odds.
II. Understand the difference between ratios, proportions and rate measures.
III. Understand in detail the definition, calculation, identification, interpretation, and application of the different measures of disease frequency (prevalence, cumulative incidence rate, incidence density rate, mortality rate).
IV. Understand the concept of person-time and when it is used (IDR vs. CIR).
V. Distinguish between case-fatality rates and mortality rates.
VI. Understand the fundamental relationships between incidence, duration and prevalence.

I. Quantifying Uncertainty - Probability and Odds

Medical data is inherently uncertain. To characterize uncertainty we can use words like "unlikely", "possible", "suspected", “consistent with”, or "probable" to describe gradations of belief. However, the exact meaning of these qualitative descriptions has been shown to vary tremendously between different clinicians. Fortunately, uncertainty can be expressed more explicitly by using *probability (*P*).* Probability expresses uncertainty on a numerical scale between 0 and 1, and is simply calculated as a *proportion* (i.e., P = *a / b* where *a* represents the number of “events” and *b* the total number at risk). Using probability avoids the ambiguity surrounding the use of qualitative descriptions – especially those like "not uncommon", or "cannot be ruled out" that too often contaminate the medical literature.

Odds are an alternative method of expressing uncertainty, and represent a concept that many people have trouble grasping. Odds represent the ratio of the probability of the event occurring over the probability of the event not occurring or P / (1 – P). For example, if the odds of an event X occurring are 1 : 9 (read as “one to nine”), this implies that the event X will occur once for every 9 times that event X will not occur. Or in 10 events X will occur once or 10% of the time.

Probability (P) and odds both quantify uncertainty and can be converted back and forth using the following formulas:

$$\text{Odds} = \frac{P}{1 - P} \quad \text{and,} \qquad P = \frac{\text{Odds}}{1 + \text{Odds}}$$

Example: The probability of diabetes in a patient is 5%, the odds of diabetes are:

$$\text{Odds} = \frac{.05}{1 - .05} = \frac{.05}{.95} = 1 : 19$$

Example: The odds of diabetes are 1 : 19, the probability of diabetes is:

$$P = \frac{1/19}{1 + 1/19} = \frac{0.05263}{1. 05263} = 0.05 \text{ or } 5\%$$

The following table shows the relationship between probability and odds. Notice that there is little difference between probability and odds when the value of probability is small i.e., <= 10%, whereas the difference is quite marked for large probability values.

Table 1. Relationship between probability and odds

Probability	Odds
0.80	4
0.67	2
0.60	1.5
0.50	1.0
0.40	0.67
0.33	0.5
0.25	0.33
0.20	0.25
0.10	0.11
0.05	0.053
0.01	0.0101

Using the equations provided above, practice converting probabilities to odds (and vice versa). The importance of being able to master this conversion will become more apparent when we discuss diagnostic (clinical) testing and Bayes' Theorem in later lectures.[1]

Note that in clinical epidemiology, probability and odds are often used to express a physician's *opinion* about the *likelihood* that an event will occur, rather than necessarily the absolute probability or odds that an event will occur. This emphasis on quantifying a physician's *opinion* will again become more evident when we come to discuss Bayes' Theorem.

[1] For those of you with more than a passing interest in betting, note that betting odds are typically odds *against* an event occurring, rather than the odds *for* an event occurring. For example, if at the track the odds for a horse to win a race are 4-1 ("four to one"), this means that 4 times out of five the horse would not be expected to win the race i.e., the probability of winning is only 20%. In the unusual race where there is a clear favourite, the odds are switched to become odds in favour of an event. This is indicated by the statement "odds on favourite". For example, the statement "Rapid Lad is the 5-4 odds on favourite for the 2.30 pm Steeplechase", means that 5 out of 9 times the horse would be expected to win (the probability is therefore 55% - which is still only about an even shot). Got it?

II. Measures of Disease Frequency

A fundamental aspect of epidemiology is to quantify or measure the occurrence of illness in a population. Obtaining a measure of the disease occurrence or impact is one of the first steps in understanding the disease under study.

Ratios, Proportions and Rates

There are three types of descriptive mathematical statistics or calculations which are used to describe or quantify disease occurrence: ratios, proportions and rates.

i. **Ratios**

A ratio is expressed as: $\frac{a}{b}$ ("a" is not part of "b")

where a and b are two mutually exclusive frequencies, that is to say the **numerator** (= a, the number on top of the expression) is not included in the **denominator** (= b, the number on the bottom of the expression).

Examples:

i) The ratio of blacks to whites in a particular school was 15/300 or 1:20. Note that the two quantities are mutually exclusive - blacks are not included as whites (and vice versa). The observed frequencies in a ratio are often re-expressed by dividing the smaller quantity into the larger one. Thus dividing 15 into 300 re-expresses the ratio in terms of 1 in 20.

ii) The ratio of spontaneous abortions to live births in a village was 12/156 or 1:13. Again note the exclusiveness of the two frequencies - abortions cannot be included as live births.

ii. **Proportion**

A proportion expresses a fraction in which the numerator (the frequency of disease or condition) is included in the denominator (population). Fractions may be multiplied by 100 to give a percentage.

Proportion = $\frac{a}{b}$ ("a" is included in "b")

Percentage = $\frac{a}{b} \times 100 = \%$

Examples:

i) The proportion of blacks in the school was 15/315 = 0.048 or 4.8%.

ii) Of 168 women that were confirmed pregnant by ultrasound examination, 12 had spontaneous abortion, thus the proportion of abortions was 12/168 = 0.071 or 7.1%.

iii. **Rates**

Rates are special types of proportions which express the relationship between an event (e.g., disease) and a defined population-at-risk evaluated over a specified time period. The numerator is the number of affected individuals in a given time period, while the denominator is the population at risk over the same time period.

$$\text{Rate} = \frac{a}{b} \qquad \begin{array}{l}\text{("a" is included in "b")} \\ \text{("b" represents population-at-risk)}\end{array}$$

The essential elements of any rate are the definition of both a *population-at-risk* and a *specific time period* of interest. As discussed below there are two types of rates commonly used as epidemiologic measures: the cumulative incidence rate and the incidence density rate.

III. Prevalence, Cumulative Incidence Rate and Incidence Density Rate

There are three basic measures of disease frequency used in epidemiology: prevalence, cumulative incidence and incidence density. These measures are commonly confused, so understanding the differences between these measures is critical.

i. Prevalence

The prevalence of disease is the *proportion* of the number of cases observed compared to the population at risk at a given point of time.

$$\text{Prevalence} = \frac{\textit{Number of cases observed at time } t}{\textit{Total number of individuals at time } t}$$

Prevalence refers to **all cases** of disease observed **at a given moment** within the group or population of interest, whereas incidence (with which it is often confused), refers to **new cases** that have occurred **during a specific time period** in the population.

Sometimes you will see a distinction made between a point prevalence and a period prevalence. The former refers to the prevalence at an exact point in time (e.g., a given day), whereas the latter refers to the prevalence during a particular time interval (e.g., during a week, month or year). For example, in a health survey we might ask the following 2 questions about the prevalence of asthma symptoms:

1. Have you had asthma symptoms today?
2. Have you had asthma symptoms any time in the last month?

Question 1 is measuring the point prevalence of asthma symptoms, while question 2 is measuring the period prevalence.

Example: Calculation of the prevalence of diarrhea on a cruise ship

You are asked to investigate an outbreak of diarrhea on a cruise ship. On the day you visit the ship, you find 86 persons on the ship. Of these you find that 8 are exhibiting signs of diarrhea. The prevalence of diarrhea at this particular time is therefore 8/86 = 0.092 or 9.2%.

Other examples:

i) The prevalence of glaucoma in a nursing home on March 24th was 4/168 = 0.024.

ii) The prevalence of smoking in Michigan adults as measured by the Behavioural Risk Factor Survey in 2002 was 26.5%.

Prevalence is a function of both the incidence rate (see below for definition of incidence) and the mean duration of the disease in the population.

Prevalence = Incidence X Duration

So, for a given incidence rate, the prevalence will be higher if the duration of the disease is longer - as an example, the prevalence of arthritis in an elderly population is high since there is no cure for the condition so once diagnosed the person has it for the rest of their lives. The prevalence will also be affected by the mortality rate of the disease, a lower prevalence would result if the disease was usually fatal – as an example, the prevalence of rabies will always be extremely low because it is almost universally fatal. Incidence rather than prevalence is usually preferred in epidemiologic studies when the objective is to convey the true magnitude of **disease risk** in the study population. Conversely, prevalence is often preferred to incidence when the objective is to convey the true magnitude of **disease burden** in the study population – particularly for chronic disease like arthritis, diabetes, mental health etc.

There are two different measures of disease incidence:

ii. <u>Cumulative Incidence Rate (Risk)</u>

The most commonly used measure of incidence – particularly in clinical studies is the cumulative incidence rate (CIR), which is also referred to as "risk". The CIR is defined as the *proportion* of a fixed population that becomes diseased during a stated period of time. Cumulative incidence incorporates the notions of a population-at-risk and a specific time period, hence it is regarded as a rate.

$$CIR = \frac{\textit{Number of newly disease individuals for a specific time period}}{\textit{Total number of population-at-risk for same time period}}$$

The CIR has a range from 0 to 1 and must be accompanied by a specified time period to have any meaningful interpretation (this is because the proportion of the population affected generally increases over time, thus it is important to know over what time period the CIR is calculated). The CIR is a measure of the ***average risk***, that is, the probability that an individual develops disease in a specified time period. For example, if a doctor tells you that the 5-year CIR of disease X is 10%, then this implies that you have a 10% *risk* of developing the disease over the next 5 years (note that the term "*risk*" may also be described as the "*chance*" or "*likelihood*" of developing the disease). Finally, note that in evidence-based medicine parlance the CIR is often referred to as the "***event rate***", and in the context of randomized trials, the CIR in the control group is called either the control event rate (CER) or the baseline risk, whilst the CIR in the treatment group is called either the experimental event rate (EER) or treatment event rate (TER). (Note that the FF text also refers to CIR as the Absolute risk in Table 5.3)

Other important CIRs:

A specific type of CIR is the **Case-Fatality Rate** (CFR) which is the proportion of affected individuals who die from the disease. In our cruise ship example, if 3 of the 8 affected people had died as a result of the diarrhea then the case-fatality rate would have been 3/8 = 0.37 or 37%. The case-fatality rate is usually associated with the seriousness and/or the virulence of the disease under study (the higher the case fatality rate the more virulent the disease). Note the important distinction between the CFR and the mortality rate – the key difference is that the denominator of the CFR is affected (diseased) subjects, whereas the denominator of the mortality rate is the whole population-at-risk (see later in the notes for a fuller discussion).

Another specific type of CIR is the **Attack Rate** which is commonly used as a measure of morbidity (illness) in outbreak investigations. It is calculated simply as the number of people affected divided by the number at risk. For our cruise ship example, after 5 days of the outbreak 12 people developed diarrhoeal disease. The attack rate at that time was therefore 12/86 = 0.14 or 14%.

iii. Incidence Density Rate

A second type of incidence rate which is more commonly used in larger epidemiologic studies is the incidence density rate (IDR). The IDR is a measure of the *instantaneous force or speed* of disease occurrence.

$$IDR = \frac{\textit{Number of newly disease individuals}}{\textit{Sum of time periods for all disease-free individuals-at-risk}}$$

The numerator of the IDR is the same as the CIR – that is, the number of newly diseased individuals that occur over time. However, it is the denominator of the IDR that is different - it now represents the sum of the disease-free time experience for all the individuals in the population. The denominator of the IDR is termed "*person-time*" or "*population time*" and represents the total disease-free time experience for the population- at-risk (the concept of person-time is explained further below). The IDR ranges from 0 to infinity, while its dimensionality is the reciprocal of time i.e., time^{-1}.

Whereas, the CIR simply represents the proportion of the population-at-risk who are affected over a specified time period, the IDR represents the *speed or instantaneous rate* at a given point in time that disease is occurring in a population. This is analogous to the speed with which a motor car is traveling - that is, miles per hour is an instantaneous rate which expresses the distance traveled for a given unit of time. An incidence rate of 25 cases per 100,000 population-years expresses the instantaneous speed which the disease is affecting the population. The IDR is a dynamic measure meaning it can change freely just as the speed of a car can. An IDR of 0 implies that the disease is not occurring in a population, whereas, an IDR of infinity is its theoretical maximum value and implies an instantaneous, universal effect on the population (if you want an example of this - think of a catastrophic event that kills everyone instantaneously – like a nuclear explosion! – this translates to a mortality rate of infinity).

The concept of person-time:

The calculation of "person-time" requires that the disease-free time contributed by each individual is summed across everyone in the population. As an example:

> Scenario A: 100 people are followed for 1 year and all remain healthy, so they contribute 100 years of disease-free person-time (i.e., 100 x 1 year).

Scenario B: 200 people are followed for 6-months and all remain healthy, so they contribute 100 years of disease-free person-time (i.e., 200 x 0.5 year).

Scenario C: 100 people are followed for 1-year, 80 remain healthy but 20 develop disease at an average of 6-months. The total disease-free person-time is now 90 years (i.e., 80 x 1 year + 20 x 0.5-year).

The particular unit of "population time" (i.e., days, weeks, months, years) that is used depends entirely on the context of the study and the disease. In chronic disease studies, a standard measure is 100,000 person years, whereas in infections disease (especially outbreak investigations we might chose to express population time in terms of person- days or person weeks - whatever makes the most sense). Since it is obviously impractical to count the exact person-time for large populations, person time is usually approximated by estimating the population size midway through the time period. For example, to calculate the mortality rate for people 60 - 65 year of age in 2000, the population time can be approximated by estimating the size of this population on July 1st, 2000 and

expressing this value in person years. An example of this is shown in the following table which shows the Lung cancer incidence and mortality rates from the 2005 Michigan cancer registry in 2005. The incidence and mortality rates are estimated using the total population that was estimated to be living in Michigan on July 1st, 2005. Obviously, in this example the exact person-time experience was not calculated for all 10 million people living in the state – such precision is unnecessary when dealing with such large populations and relatively rare events. Hint: Be sure to confirm the calculation of these rates in the table below.

Lung Cancer Incidence and Mortality. Michigan Cancer Registry 2005:

	Michigan 2005 Population Estimated on July 1st 2005	Number of new lung cancer cases diagnosed in 2005	Number of deaths due to Lung cancer in 2005	Incidence rate of Lung Cancer per 100,000 person years	Mortality rate from Lung Cancer per 100,000 person years
Total	10,125,000	7,681	5,789	75.8	57.1
Male	4,975,000	4,218	3,179	84.8	63.9
Female	5,150,000	3,463	2,610	67.2	50.7

The choice between CIR or IDR?

A natural question to ask is when should the CIR be the incidence measure of choice, and when should the IDR be preferred? What follows are some general guidelines, although there are no hard and fast rules.

The CIR tends to be used when there is a **fixed or closed** population that is starting at a common point in time. The enrollment of a medical school class is a good example of this - a fixed group of students start on the same day and the class is closed to further admissions (and hopefully deletions). Also, the CIR only counts the first event in a given individual, which may be just fine if the health event or outcome can only occur once (e.g., death), but may not satisfactorily capture the true picture for outcomes that can have multiple events – such as infections, or hospitalizations or births.

In contrast, the IDR is used when there is an **open** population - subjects can move in and out of the population, or when the starting point is not fixed. A good example of this is a randomized trial where it may take many months to enroll enough patients – thus the amount of follow-up time of each patient will vary depending on when subjects were enrolled. The IDR is also the preferred measure when the outcome can occur more than once within an individual e.g., upper respiratory tract (URT) infection.

The CIR is more suitable for measuring disease event rates that are relatively stable over time (such as cancer statistics), whereas the IDR is more useful when the disease event rates are highly variable such as occurs in outbreaks of infectious diseases. The following table summarizes the use of the two measures (again there are no hard and fast rules in this so you will see examples that don't follow the table below):

	CIR	**IDR**
Population	Closed	Open
Starting Point	Fixed	Variable
Type of outcome	Single (death)	Single or Multiple (URT infection)
Fluctuation in underlying event rates	Stable (cancer stats)	Highly variable (outbreaks)
Example	Fixed cohort (medical school class)	Open cohort (RCT)

The Mortality Rate:

The mortality rate describes the frequency of death in a defined population during a specified time period. We can measure the mortality rate using either a cumulative incidence rate (CIR) or an incidence density rate (IDR) – the only thing that distinguishes a mortality rate from an incidence rate is that we are measuring the number of deaths in the population rather than the number of disease events.

$$\textit{Mortality rate (CIR)} = \frac{\textit{Number of deaths over a specific time period t}}{\textit{Total number of population-at-risk for same time period t}}$$

$$\textit{Mortality rate (IDR)} = \frac{\textit{Number of deaths over a specific time period t}}{\textit{Sum of time periods for all individuals-at-risk}}$$

The denominator for the mortality rate is the size of the population who are at risk of dying from the condition – which is either specified as the total number (in the CIR) or as the total population time (in the IDR).

When calculated for all deaths combined (regardless of cause) the mortality rate is referred to as *all-cause mortality*. Alternatively, if the rate of death is calculated for a specific cause (e.g., lung cancer or prostate cancer) the rate is referred to as *disease-specific mortality*.

Note that without disease incidence we cannot have disease mortality – the biggest contributor to mortality is incidence. The mortality rate is therefore some fraction of the underlying incidence rate depending on the lethality of the condition. For example, for lethal conditions such as lung cancer the mortality rate is very close to the underlying incidence rate, whereas for more benign conditions such as prostate cancer, the mortality rate is a much smaller fraction of the underlying incidence. The lethality of the condition is best captured by the case fatality rate (or survival rate), as illustrated in the following table that shows U.S. cancer statistics from 2003-2007 (based on the SEER Registry) (http://seer.cancer.gov/csr/1975_2007/results_merged/topic_survival.pdf):

Population (Cancer site)	Age adjusted Incidence per 100,000	Age adjusted Mortality per 100,000	5-year relative survival (%) 1999-2006
Total (Lung Cancer)	62.5	52.5	15.8
Men (Lung Cancer)	76.2	68.8	13.5
Women (Lung Cancer)	52.4	40.6	18.3
Men (Prostate Cancer)	150.4	22.8	99.6
Women (Breast Cancer)	126.5	23.4	90.2

Clearly the mortality rate should be distinguished from the case-fatality rate (or survival rate). The denominator for the case-fatality rate is the number of affected individuals, not the population at risk. The use of these 2 terms are often confused, yet they are very different as illustrated in the above table and in the following example:

A study by McGovern (NEJM, 1996;334:884-90) looked at trends in mortality and survival from acute myocardial infarction (MI) in the general population of Minneapolis, MN. They found the 28-day acute MI CFR for men and women to be 10% and 12%, respectively, while the mortality rates for acute MI in men and women were 110 per 100,000 person years, and 35 per 100,000 person years, respectively. So while the CFRs were very similar, the mortality rates were very different. Only the mortality rate confirms what your intuition tells you - that the mortality from MI is higher in men than women – in this case the death rate from acute MI is about 3 times higher. Note that this difference is “driven” by a higher incidence rate of MI in men compared to women (and not case fatality).

Optional Section.

Finally, for the mathematicians among you, the following section explains how the CIR and IDR are related to each other

Optional section:

What is the relationship between the CIR and the IDR?

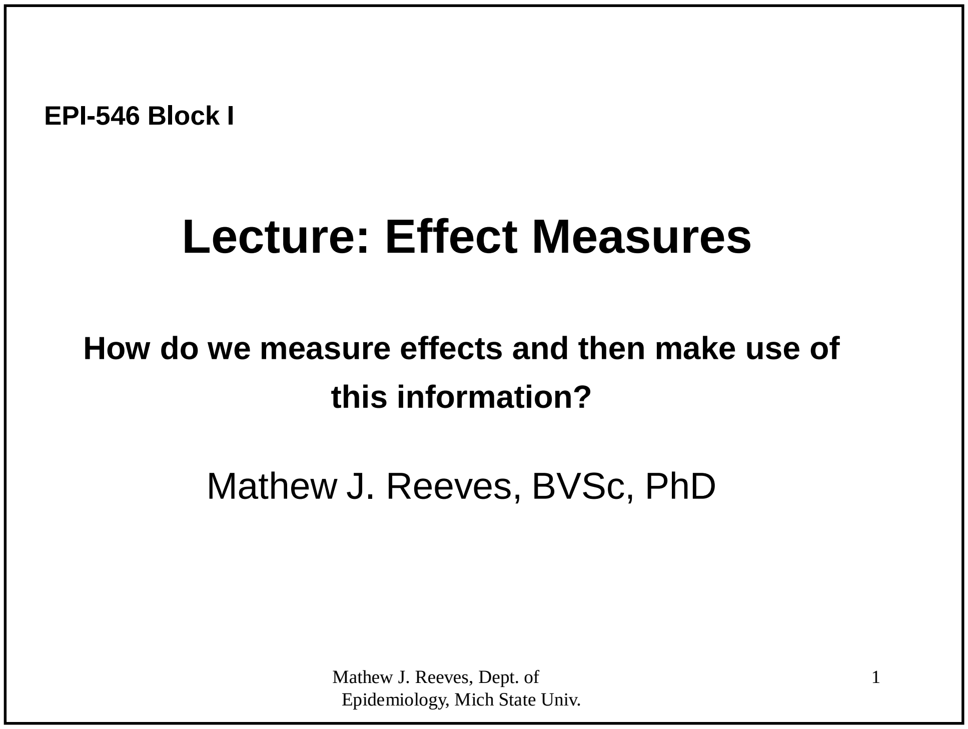
Assuming a constant incidence rate (IDR) and no other causes of disease (i.e., competing risks). The relationship between the CIR and IDR is described by the following exponential function:

- $CIR_t = 1 – e^{– IDR\ \Delta\ t}$ where t = time

When the IDR is small (< 0.1), then

- $CIRt \sim = IDR\ \Delta\ t$

So to estimate small risks, simply multiple the IDR by the time period. For example, if the IDR = 10 / 1,000 PY (equivalent to 1%), the 5-year CIR or risk is 5 x (10/1,000) or 5%.

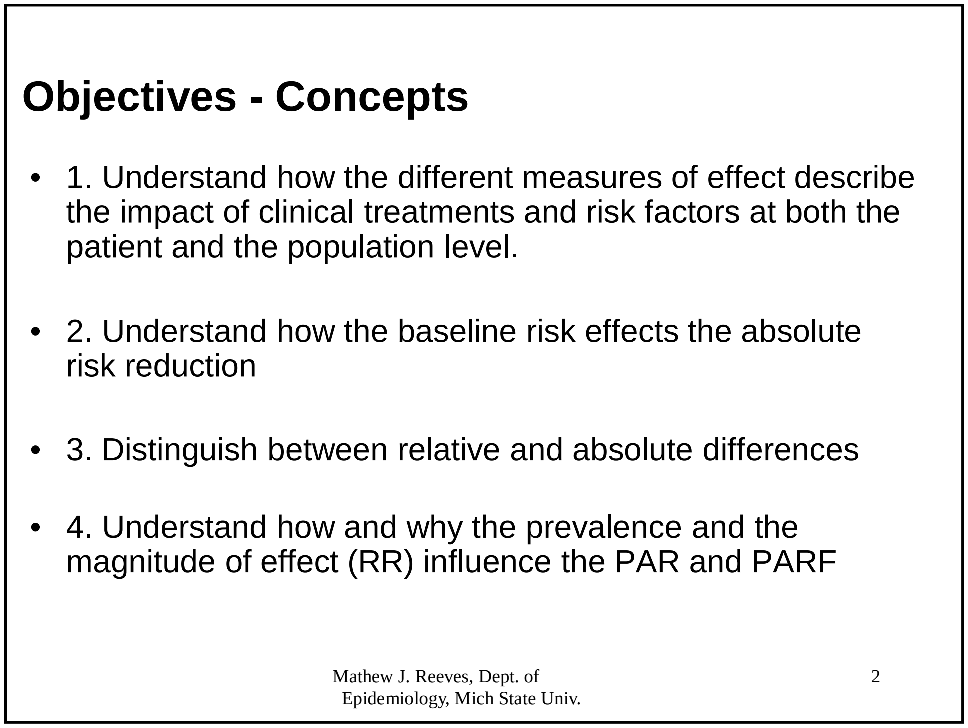


EPI-546 Block I

Lecture: Effect Measures

How do we measure effects and then make use of this information?

Mathew J. Reeves, BVSc, PhD



Objectives - Concepts

- 1. Understand how the different measures of effect describe the impact of clinical treatments and risk factors at both the patient and the population level.
- 2. Understand how the baseline risk effects the absolute risk reduction
- 3. Distinguish between relative and absolute differences
- 4. Understand how and why the prevalence and the magnitude of effect (RR) influence the PAR and PARF

Objectives - Skills

- 1. Define, calculate, identify, interpret and apply measures of effect (RR, RRR, AR, ARR, ARI, NNT, NNH, PAR, PARF)
- 2. Understand how to calculate the OR and know when it is a good approximation to the RR

Measures of Effect

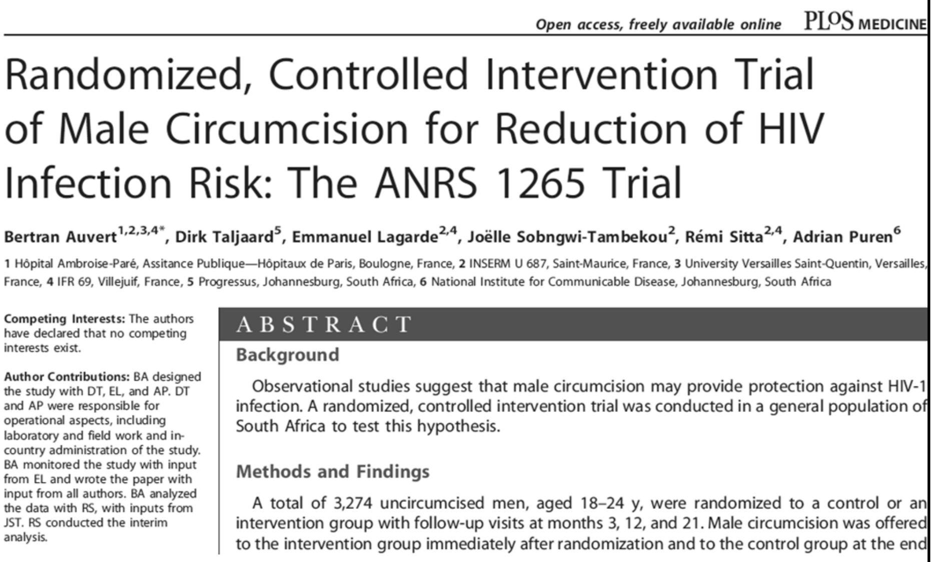
- Presentation and Interpretation of Information on Risk

- Information on the effect of a treatment or risk factor can be presented in several different ways
  - Relative Risk (RR)
  - Relative Risk Reduction (RRR)
  - Absolute Risk Reduction (ARR)
  - Absolute Risk Increase (ARI)
  - NNT (Number needed to treat)
  - NNH (Number needed to harm)
  - Population attributable risk (PAR)
  - Population attributable risk fraction (PARF)
  - The Odds Ratio (OR)
- The way risk information is presented can have a profound effect on clinical decisions (both on part of patients and doctors)

The 2 x 2 Table – Clinical Intervention Study (RCT)

		Outcome		
		Yes	**No**	
Treatment Group	**Intervention (t)**	a	b	Risk$_t$ = a / a + b
	Control (placebo) (c)	c	d	Risk$_c$ = c / c + d

(Risk = CIR)



Open access, freely available online PLoS MEDICINE

Randomized, Controlled Intervention Trial of Male Circumcision for Reduction of HIV Infection Risk: The ANRS 1265 Trial

Bertran Auvert[1,2,3,4*], Dirk Taljaard[5], Emmanuel Lagarde[2,4], Joëlle Sobngwi-Tambekou[2], Rémi Sitta[2,4], Adrian Puren[6]

1 Hôpital Ambroise-Paré, Assitance Publique—Hôpitaux de Paris, Boulogne, France, 2 INSERM U 687, Saint-Maurice, France, 3 University Versailles Saint-Quentin, Versailles, France, 4 IFR 69, Villejuif, France, 5 Progressus, Johannesburg, South Africa, 6 National Institute for Communicable Disease, Johannesburg, South Africa

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: BA designed the study with DT, EL, and AP. DT and AP were responsible for operational aspects, including laboratory and field work and in-country administration of the study. BA monitored the study with input from EL and wrote the paper with input from all authors. BA analyzed the data with RS, with inputs from JST. RS conducted the interim analysis.

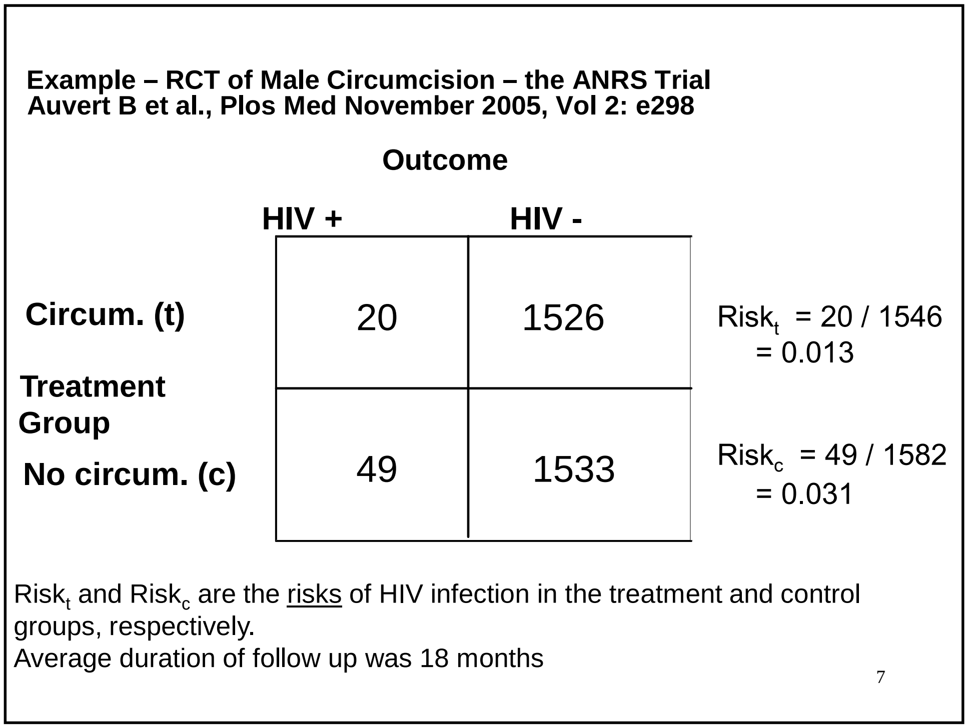
ABSTRACT

Background

Observational studies suggest that male circumcision may provide protection against HIV-1 infection. A randomized, controlled intervention trial was conducted in a general population of South Africa to test this hypothesis.

Methods and Findings

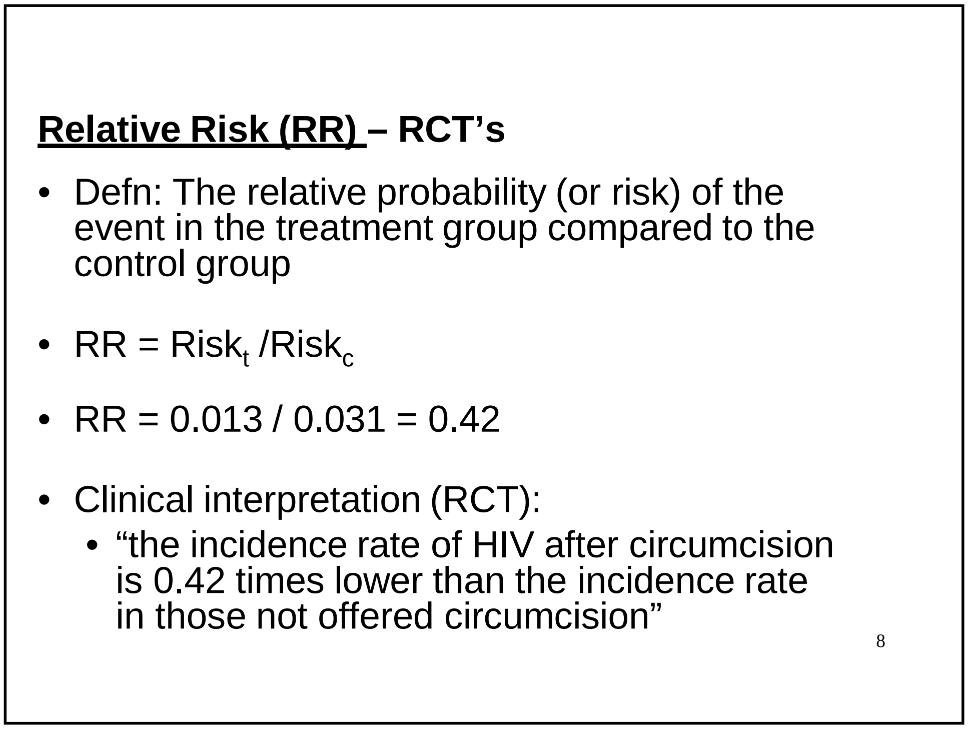
A total of 3,274 uncircumcised men, aged 18–24 y, were randomized to a control or an intervention group with follow-up visits at months 3, 12, and 21. Male circumcision was offered to the intervention group immediately after randomization and to the control group at the end



Example – RCT of Male Circumcision – the ANRS Trial
Auvert B et al., Plos Med November 2005, Vol 2: e298

		Outcome		
		HIV +	HIV -	
Treatment Group	Circum. (t)	20	1526	$Risk_t$ = 20 / 1546 = 0.013
	No circum. (c)	49	1533	$Risk_c$ = 49 / 1582 = 0.031

$Risk_t$ and $Risk_c$ are the risks of HIV infection in the treatment and control groups, respectively.
Average duration of follow up was 18 months



Relative Risk (RR) – RCT's

- Defn: The relative probability (or risk) of the event in the treatment group compared to the control group
- RR = $Risk_t$ /$Risk_c$
- RR = 0.013 / 0.031 = 0.42
- Clinical interpretation (RCT):
  - "the incidence rate of HIV after circumcision is 0.42 times lower than the incidence rate in those not offered circumcision"

<u>Relative Risk (RR)</u> – RCT's

- A measure of the <u>efficacy</u> of a treatment
- Null value = 1.0.
- RR < 1.0 = decreased risk (beneficial treatment)
- RR > 1.0 = increased risk (harmful treatment)
- Not a very useful measure of the clinical impact of treatment (need ARR)

9

<u>Relative Risk Reduction (RRR)</u>

- Defn: The proportion of the baseline risk that is removed by therapy
- RRR = 1 – RR
- RRR = 1 - 0.42 = 0.58 or 58%
- Clinical interpretation (RCT):
  - "the HIV infection rate is 58% lower after circumcision compared to no circumcision"

10

Relative Risk Reduction (RRR)

- Indicates by how much in relative terms the event rate is decreased.
- Used to quantify the effect of clinical treatments - “How much does this treatment reduce the disease or outcome?”
- Can also calculated as the ARR divided by the baseline risk
  - ARR/ Risk_c = [0.031-0.013]/0.031 = 58%
- Null value = 0.

RR/RRR – RCT – Interpretation?

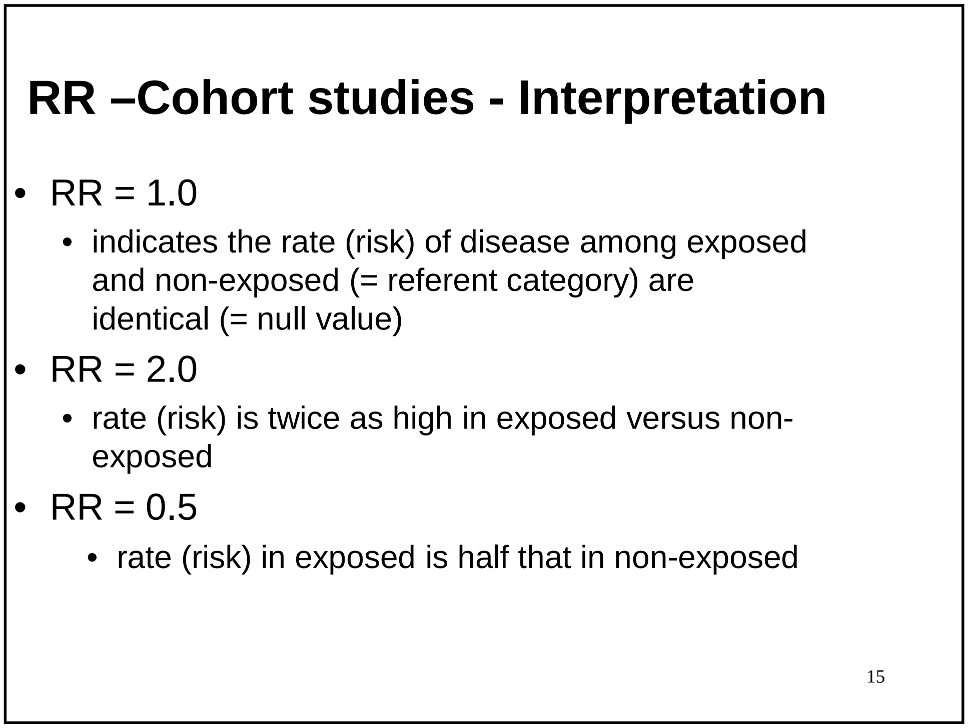
- RR = < 0.5 or > 2.0
  - **BIG**
  - RRR= 50% or RRI= 100% (doubling)
- RR = 0.5 - 0.8 or 1.25 – 2.0
  - **MODERATE to BIG**
  - RRR= 20-50% or RRI= 25-100%
- RR = ~ 0.90 or ~ 1.10
  - **SMALL**
  - 10% reduction or 10% increase

<u>Relative Risk (RR)</u> – Cohort studies

- In cohort studies, RR is also used to measure the <u>magnitude</u> of association between an exposure (risk factor) and an outcome (See study design lectures).
- Defn: The relative probability (or risk) of disease in the **<u>exposed group</u>** compared to the **<u>non-exposed group</u>**
- Example: Smoking and Lung CA
  - Lung CA mortality heavy smokers = 4.17 per 1,000 person yrs
  - Lung CA mortality non-smokers = 0.17 per 1,000 person yrs
  - RR = Risk_{exp} /$\text{Risk}_{\text{unexp}}$ = 4.17/0.17 = 24.5

<u>Relative Risk (RR)</u> – Cohort studies

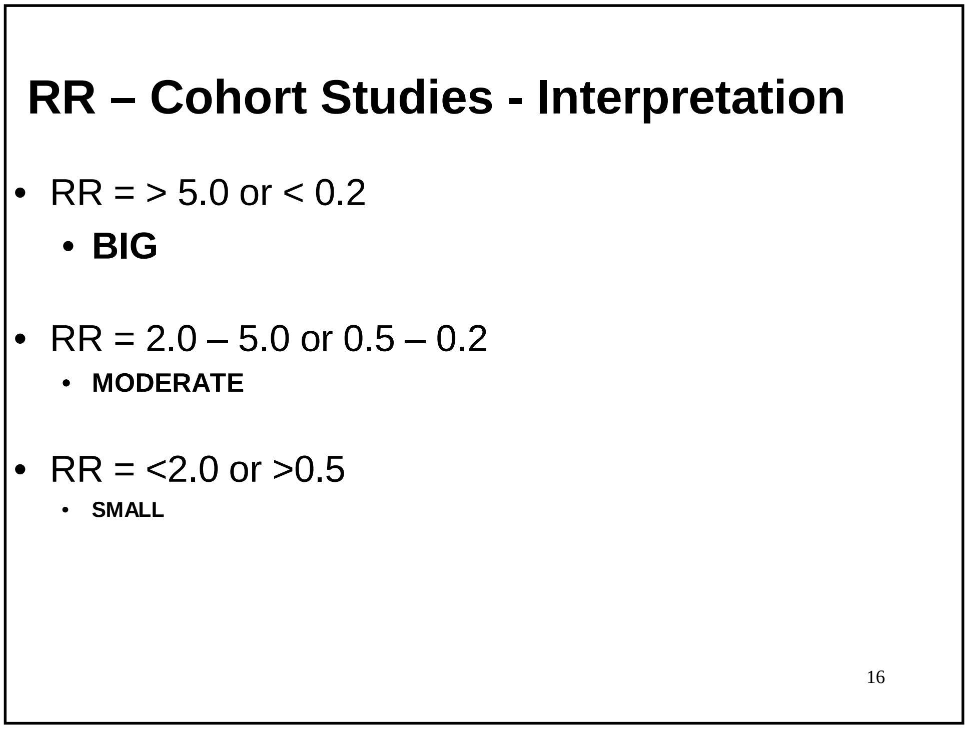
- Clinical interpretation (cohort):
  - "the risk of dying of lung CA is about 25 times higher in lifetime heavy smokers compared to lifetime non-smokers"
- To measure the impact of the risk factor in the population as a whole we also need to know the prevalence of the exposure (see PAR, PARF).



RR –Cohort studies - Interpretation

- RR = 1.0
  - indicates the rate (risk) of disease among exposed and non-exposed (= referent category) are identical (= null value)
- RR = 2.0
  - rate (risk) is twice as high in exposed versus non-exposed
- RR = 0.5
  - rate (risk) in exposed is half that in non-exposed

15



RR – Cohort Studies - Interpretation

- RR = > 5.0 or < 0.2
  - **BIG**
- RR = 2.0 – 5.0 or 0.5 – 0.2
  - **MODERATE**
- RR = <2.0 or >0.5
  - **SMALL**

16

Absolute Risk Reduction (ARR)

- Defn: The difference in absolute risk (or probability of events) between the control and treatment groups

- ARR = Risk_c - Risk_t
- ARR = 0.031 - 0.013 = 0.018 or 1.8%

- Clinical interpretation (RCT):
  - "over 18 months of follow-up the absolute risk of HIV infection was 1.8% lower with circumcision compared to no treatment"

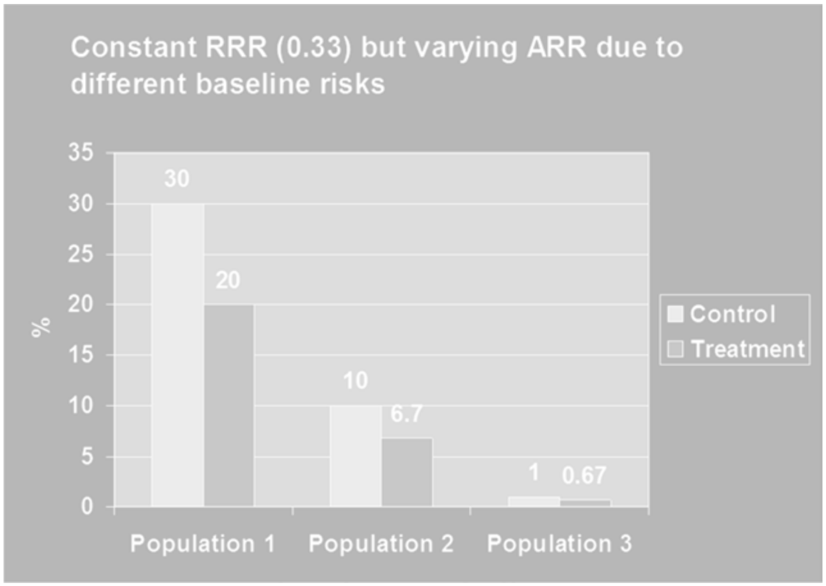
Absolute Risk Reduction (ARR)

- A simple and direct measure of the impact of treatment

- May also be called the risk difference (RD) or attributable risk

- Null value = 0.

- The ARR depends on the background baseline risk which can vary markedly from one population to another.



ARR= 10% **ARR= 3.3%** **ARR = 0.33%**

19

The Number Needed To Treat (NNT)

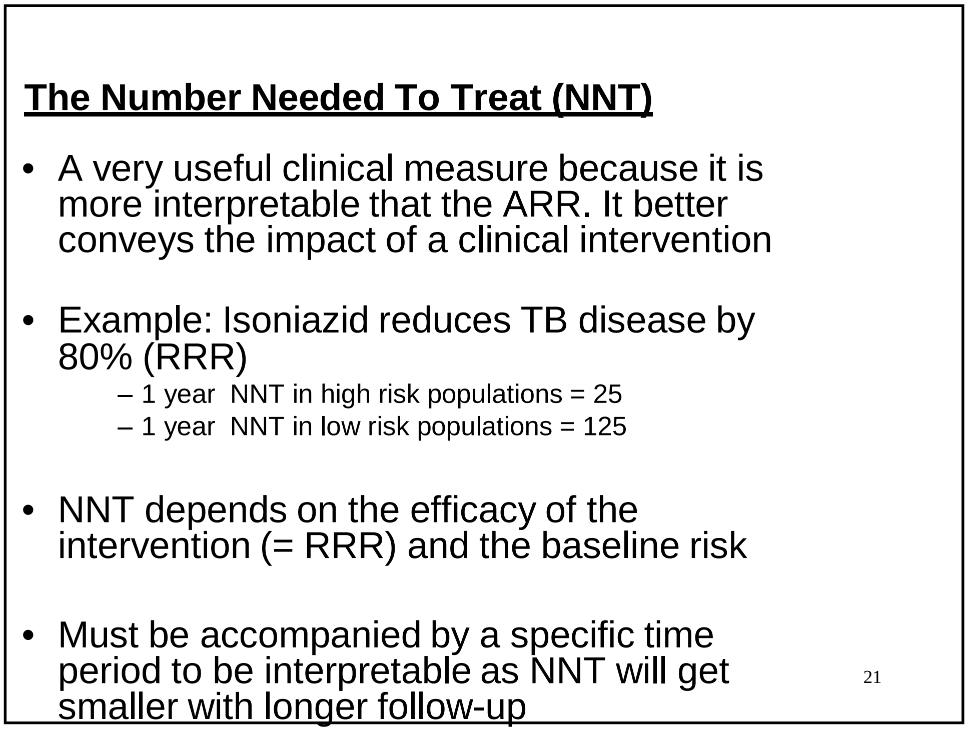
- Defn: The number of patients who would need to be treated to prevent an adverse event over a specific period of time

- NNT = 1 / ARR
  NNT = 1 / 0.018 = 55.5 (or 56)

- Clinical interpretation (RCT):
  - “over the 18 months of the study, for every 56 patients who received circumcision, one HIV infection was prevented”

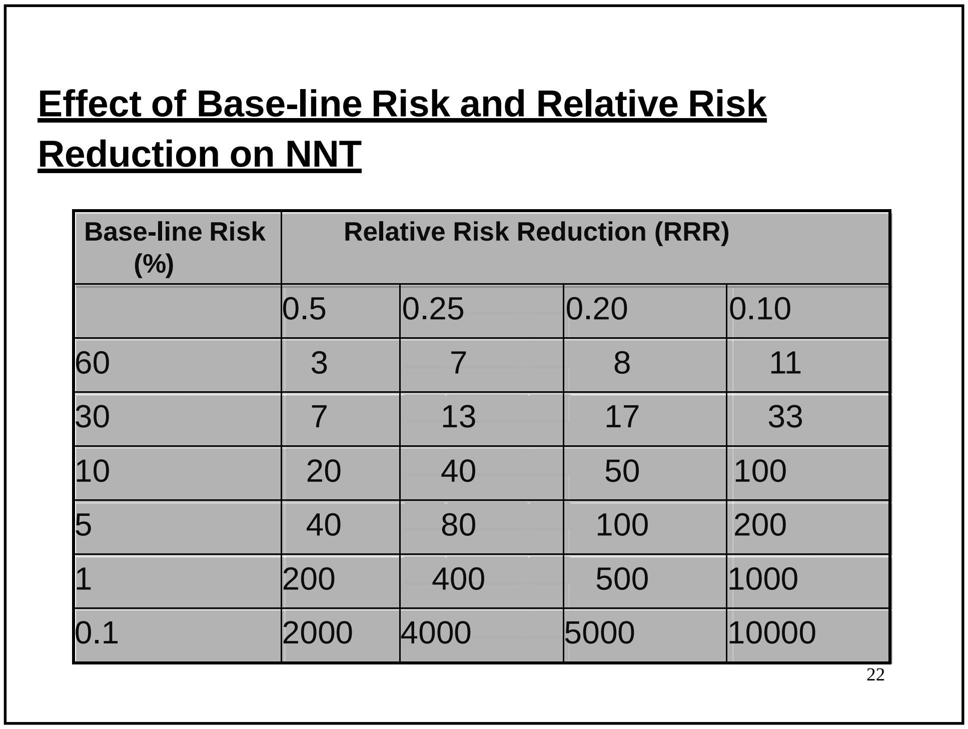
- High NNT = bad, Low NNT = good

20



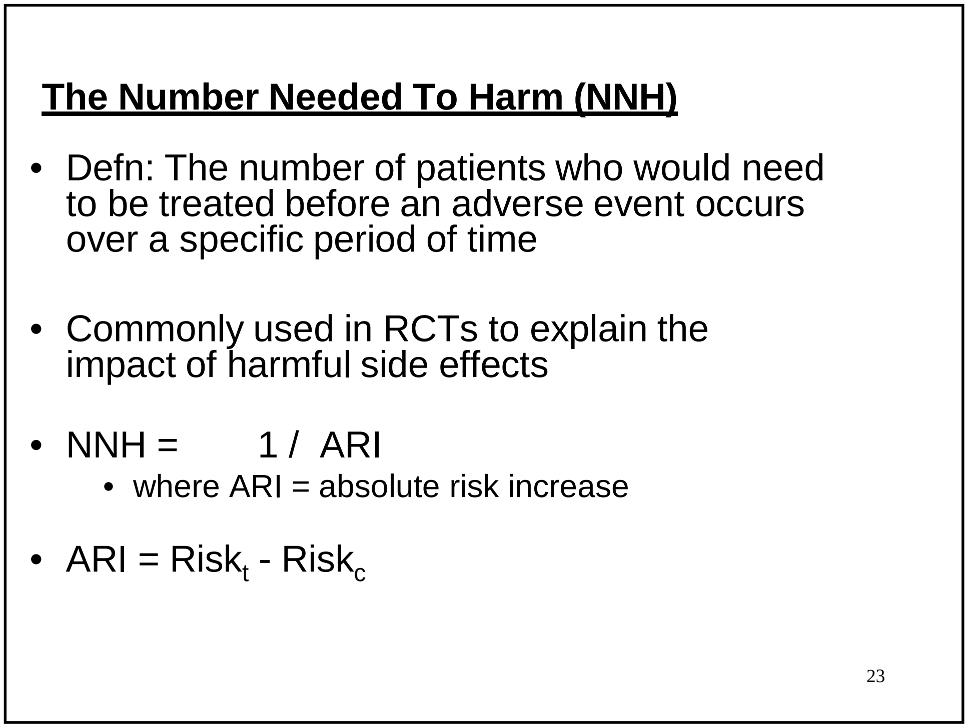
The Number Needed To Treat (NNT)

- A very useful clinical measure because it is more interpretable that the ARR. It better conveys the impact of a clinical intervention
- Example: Isoniazid reduces TB disease by 80% (RRR)
  - – 1 year NNT in high risk populations = 25
  - – 1 year NNT in low risk populations = 125
- NNT depends on the efficacy of the intervention (= RRR) and the baseline risk
- Must be accompanied by a specific time period to be interpretable as NNT will get smaller with longer follow-up



Effect of Base-line Risk and Relative Risk Reduction on NNT

Base-line Risk (%)	Relative Risk Reduction (RRR)			
	0.5	0.25	0.20	0.10
60	3	7	8	11
30	7	13	17	33
10	20	40	50	100
5	40	80	100	200
1	200	400	500	1000
0.1	2000	4000	5000	10000



The Number Needed To Harm (NNH)

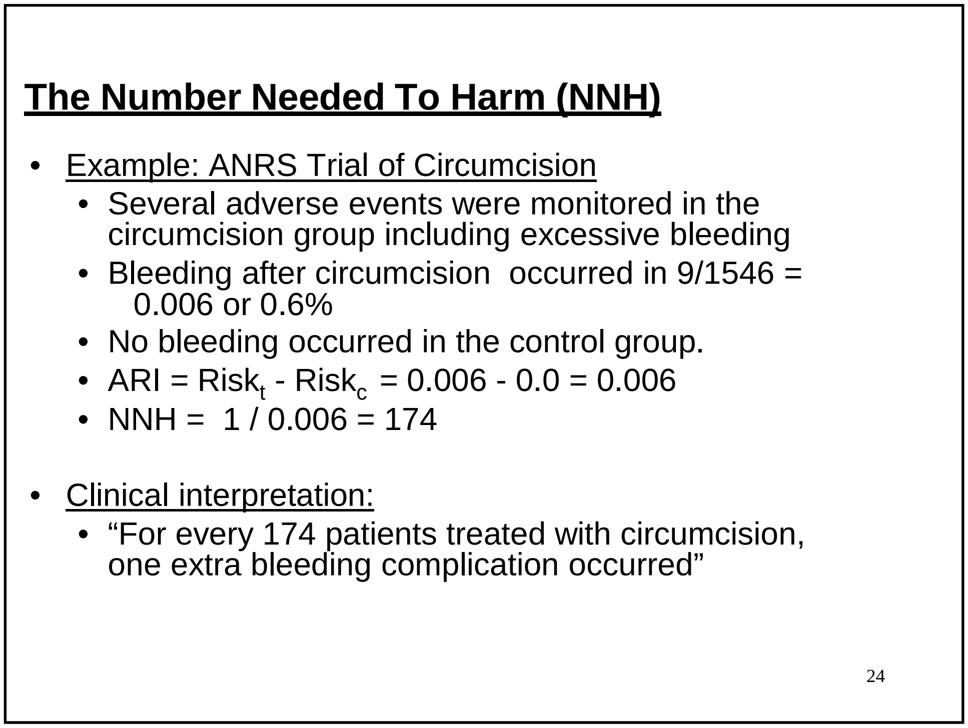
- Defn: The number of patients who would need to be treated before an adverse event occurs over a specific period of time

- Commonly used in RCTs to explain the impact of harmful side effects

- NNH = 1 / ARI
  - where ARI = absolute risk increase

- ARI = Risk_t - Risk_c

23



The Number Needed To Harm (NNH)

- Example: ANRS Trial of Circumcision
  - Several adverse events were monitored in the circumcision group including excessive bleeding
  - Bleeding after circumcision occurred in 9/1546 = 0.006 or 0.6%
  - No bleeding occurred in the control group.
  - ARI = Risk_t - Risk_c = 0.006 - 0.0 = 0.006
  - NNH = 1 / 0.006 = 174

- Clinical interpretation:
  - “For every 174 patients treated with circumcision, one extra bleeding complication occurred”

24

The Number Needed To Harm (NNH)

- Example: CHARISMA trial
  - Randomized RCT of [Clopidogrel] vs. [Placebo]
  - Mean follow-up 2.3 years
  - Safety end point: severe or fatal hemorrhage
  - Bleeding occurred in 1.7% of [Clopidgrel] group and 1.3% of [Placebo] group.
  - ARI = Risk_t - Risk_c = 0.017 - 0.013 = 0.004
  - NNH = 1 / 0.004 = 250

- Clinical interpretation:
  - "For every 250 patients treated with Clopidogrel over 2.3 years, one extra severe or fatal bleed will occur compared to placebo alone "
  - "How many patients do I need to treat to cause one bad event?"

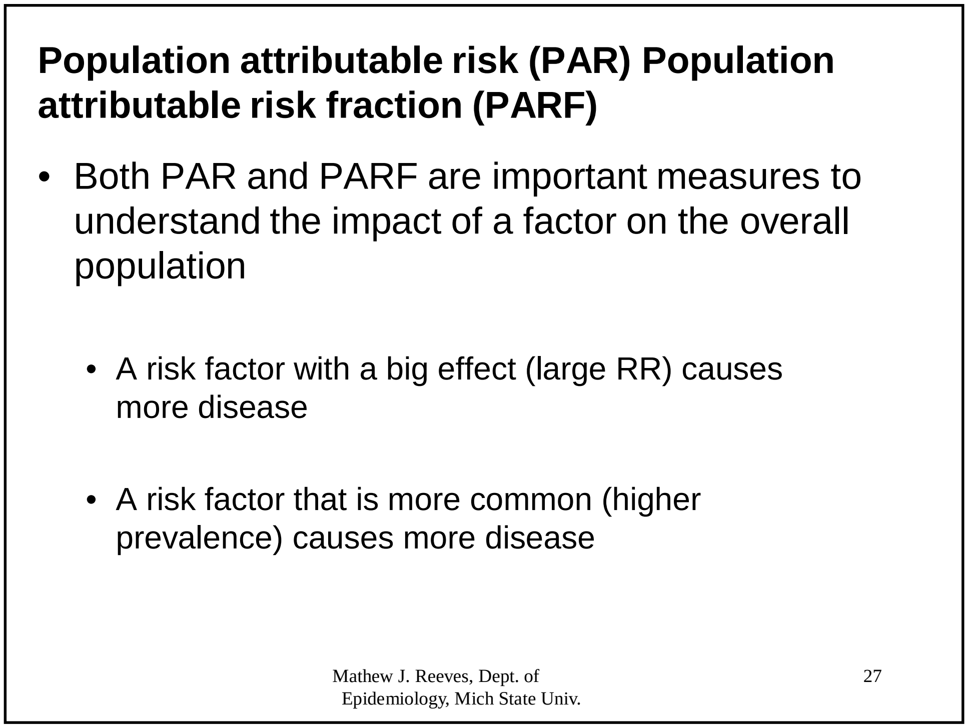
- High NNH = good, Low NNH = bad
- Again need to know the time period.

How information is conveyed (RRR, ARR or NNT) makes a difference!

- Drug effects are perceived to be much more favourable when they are presented as RRRs rather than ARRs

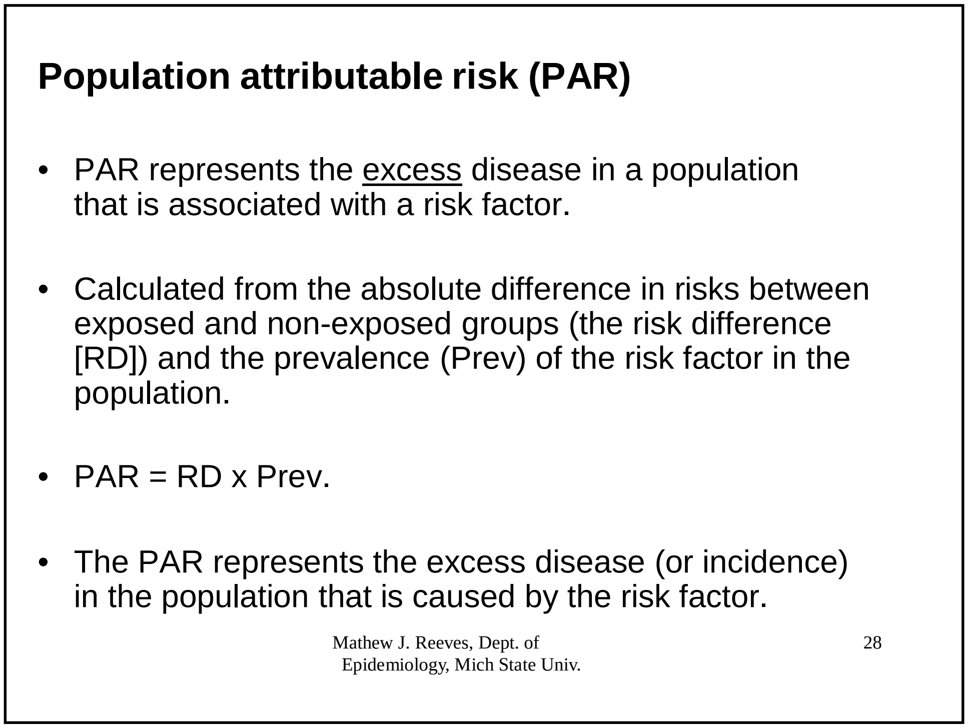
- See article by Skolbekken in the course pack.

- Pay attention to how data are 'framed' in drug advertisements.
  - **Statins reduce risk of heart attack by 40%!!!!**
  - But absolute risk reduction is only 0.5% (NNT = 200)



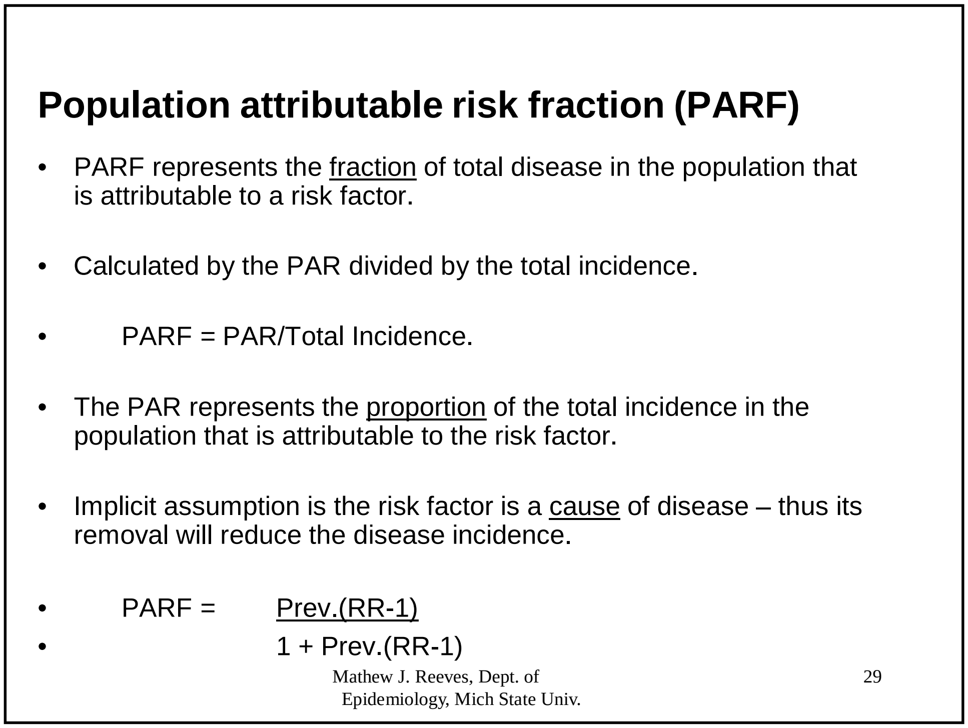
Population attributable risk (PAR) Population attributable risk fraction (PARF)

- Both PAR and PARF are important measures to understand the impact of a factor on the overall population
  - A risk factor with a big effect (large RR) causes more disease
  - A risk factor that is more common (higher prevalence) causes more disease



Population attributable risk (PAR)

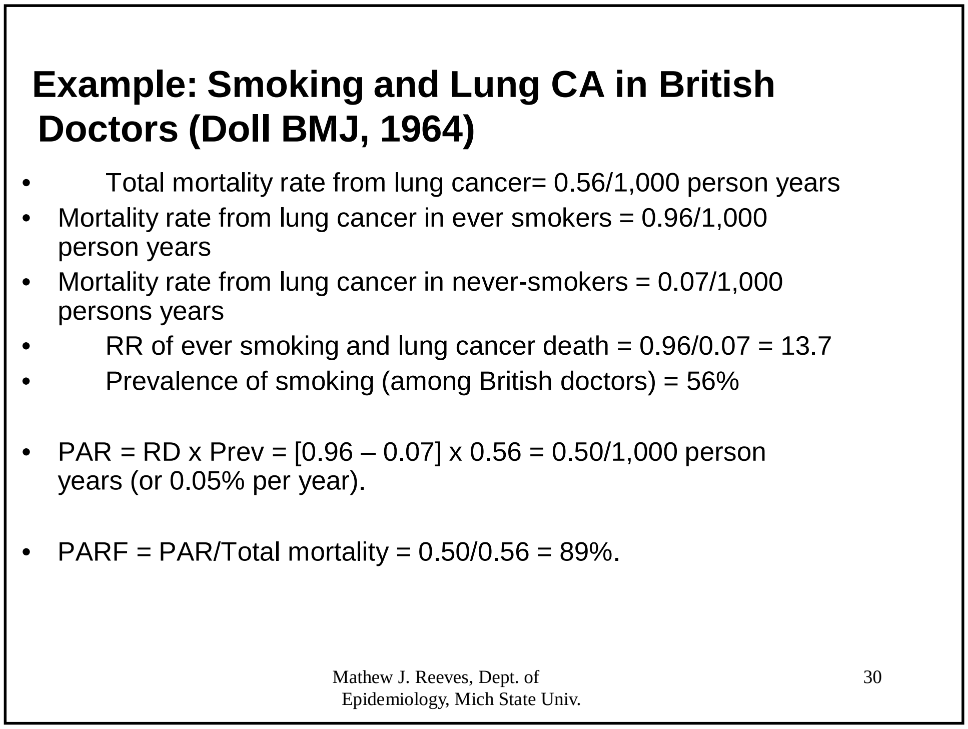
- PAR represents the excess disease in a population that is associated with a risk factor.
- Calculated from the absolute difference in risks between exposed and non-exposed groups (the risk difference [RD]) and the prevalence (Prev) of the risk factor in the population.
- PAR = RD x Prev.
- The PAR represents the excess disease (or incidence) in the population that is caused by the risk factor.



Population attributable risk fraction (PARF)

- PARF represents the fraction of total disease in the population that is attributable to a risk factor.
- Calculated by the PAR divided by the total incidence.
- PARF = PAR/Total Incidence.
- The PAR represents the proportion of the total incidence in the population that is attributable to the risk factor.
- Implicit assumption is the risk factor is a cause of disease – thus its removal will reduce the disease incidence.
- PARF = $\underline{\text{Prev.(RR-1)}}$
- 1 + Prev.(RR-1)

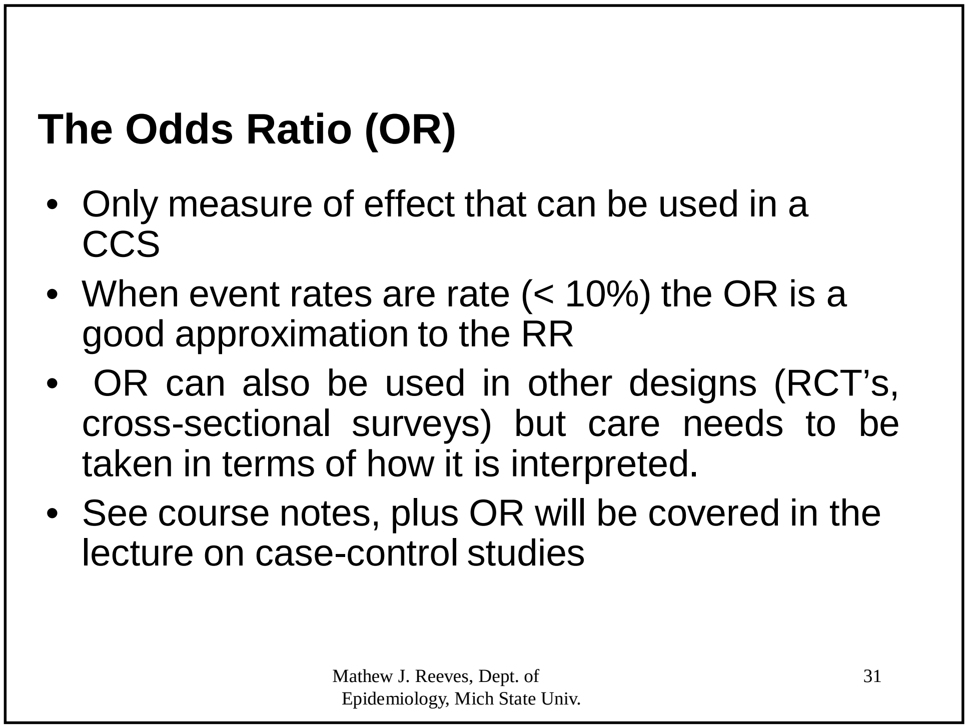
Mathew J. Reeves, Dept. of Epidemiology, Mich State Univ. 29



Example: Smoking and Lung CA in British Doctors (Doll BMJ, 1964)

- Total mortality rate from lung cancer= 0.56/1,000 person years
- Mortality rate from lung cancer in ever smokers = 0.96/1,000 person years
- Mortality rate from lung cancer in never-smokers = 0.07/1,000 persons years
- RR of ever smoking and lung cancer death = 0.96/0.07 = 13.7
- Prevalence of smoking (among British doctors) = 56%
- PAR = RD x Prev = [0.96 – 0.07] x 0.56 = 0.50/1,000 person years (or 0.05% per year).
- PARF = PAR/Total mortality = 0.50/0.56 = 89%.

Mathew J. Reeves, Dept. of Epidemiology, Mich State Univ. 30



The Odds Ratio (OR)

- Only measure of effect that can be used in a CCS
- When event rates are rate (< 10%) the OR is a good approximation to the RR
- OR can also be used in other designs (RCT's, cross-sectional surveys) but care needs to be taken in terms of how it is interpreted.
- See course notes, plus OR will be covered in the lecture on case-control studies

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes – Effect Measures

Mat Reeves BVSc, PhD

Objectives:

I. Understand how the different measures of effect are used to describe the impact of clinical treatments (in trials) and risk factors (in observational studies) at both the patient and population level.
II. Understand the definition, calculation, identification, interpretation, and application of measures of effect at the patient level (RR, RRR, AR, ARR, ARI, NNT, NNH).
III. Understand how the baseline risk affects the absolute risk reduction.
IV. Distinguish between relative and absolute differences, and understand how the use of these 2 measures can appear to imply different effects when applied to the same data.
V. Understand the definition, calculation, identification, interpretation, and application of measures of effect at the population level (PAR, PARF).
VI. The Odds Ratio (OR) – its calculation, and when is it a good approximation of the RR.

I. Risks and Measures of Effect

In clinical studies it is common to calculate the risk or CIR of an event in different populations or groups. For example, in a randomized clinical trial (RCT) the risk or CIR of an event in the treated and control groups are calculated and compared (note that these risks may also be referred to as *event rates*). By taking either the *ratio* of these two measures or the *difference* in these two measures we can calculate two fundamental *measures of effect* - the ***relative risk*** and the ***absolute risk*** – that, in the case of an RCT, quantify the impact of the treatment.

To illustrate the calculation and interpretation of these measures, we will use the following data from an RCT designed to measure the mortality rate associated with two treatments (ligation and sclerotherapy) used for the treatment of bleeding oesophageal varices (Ref: Stiegmann et al, NEJM 1992;326:1527-32). The 65 patients treated with sclerotherapy are regarded as the control group (since that was the standard of care at the time), and the 64 treated with ligation were regarded as the "new" treatment group. The risk (or CIR) of mortality was calculated for each group:

Example – RCT of Endoscopic Ligation vs. Endoscopic Sclerotherapy

	Outcome		
	Death	Survival	
Ligation (t)	18	46	$\text{Risk}_t = 18 / 64$ $= 0.28$
Sclerotherapy	29	36	$\text{Risk}_c = 29 / 65$ $= 0.45$

Where: Risk_t and Risk_c represent the risk of death in the treatment and control groups, respectively. Risk_c is often referred to as the baseline risk.

i. Relative Risk (RR)

In this RCT example, the Relative Risk (RR) is the ratio of the risk in the treated group (Risk_t or TER), relative to the risk in the control group (Risk_c or CER).

$$RR = \frac{\textit{Risk in treatment group}}{\textit{Risk in control group}} \quad \textit{or} \quad \frac{\text{Risk}_t}{\text{Risk}_c}$$

The RR is a measure of the *strength* or *magnitude* of the effect of the new treatment on mortality, *relative* to the effect of the standard (control) treatment. In this example, the RR is 0.28/0.45 = 0.62, indicating that the risk of death after ligation is 0.62 times lower than the risk of death in the sclerotherapy treated group. In other words, the risk of death with ligation is 62% (or about 2/3^{rds}) that of sclerotherapy (so if you needed treatment for oesphageal varices you'd probably pick treatment with ligation over sclerotherapy). Like all ratio measures the null value of the RR is 1.0 (i.e., the point estimate that indicates no increase or decrease in risk).

Although the RR is a very important measure of effect, it is somewhat limited in its clinical usefulness, since it fails to convey information on the likely effectiveness of clinical intervention - a better measure of the *absolute benefit* of intervention is given by the difference in risks between the two groups (i.e., absolute risk reduction (ARR)).

In the context of an epidemiologic study, specifically cohort studies, the RR measures the *strength* or *magnitude* of association between an exposure (or *risk factor)* and a disease or other outcome. It is calculated as the ratio of the risk in a group exposed to the risk factor, relative to the risk in an unexposed group. For example if the lung cancer mortality rate in lifetime heavy smokers (> 25 cigarettes/day) is 4.17 per 1,000 person years (which is equivalent to a CIR of 0.417% per year) and in lifetime non-smokers it is 0.17 per 1,000 person years (equivalent to a CIR of 0.017% per year), then the RR = 0.417/0.017 = 24.5, indicating that heavy smokers are almost 25 times more likely to die of lung cancer than non-smokers. One of the reasons that RR are favoured by epidemiologists is that they are in general, fairly constant across different populations, and so can be "transported" from one study to another (so, for example, the RR of 25 for lung cancer mortality due to heavy smoking was calculated from the famous British Doctors Study (Doll R et al, BMJ June 26th, 2004), it is likely that the effect of heavy smoking i.e., the RR is similar in the US or elsewhere).

A limitation of the RR calculated in cohort studies is that it is not a very useful measure of the *impact* of a risk factor on a population, since it does not include any information on the frequency or prevalence of the risk factor (e.g., smoking) in the population (see Population Attributable Risk Fraction).

Note that the Odds Ratio (OR) (see below) has a similar interpretation as the RR in that it also measures the strength or magnitude of the association between exposure and outcome.

An important point about the RR is that it is only calculated in study designs where the actual incidence or risk of an event is measured i.e., RCTs and cohort studies. The interpretation of the magnitude a particular RR depends on the type of study it was generated from. For cohort studies, a general rule for a factor that increases risk of an outcome is that RR values of 2.0 or less are regarded as *small*, values of >2 to 5 are *moderate*, and those >5 are *large* effects. For a factor that decreases risk the equivalent RR estimates for small, medium and large effects are >0.5, 0.5-0.2, and <0.2, respectively. These same rules also apply to the OR. The magnitude of the RR is important to epidemiologists because the larger the value the less likely it is that the particular relationship is due to chance or bias (such as confounding or selection). For example, one of the reasons that smoking is regarded as cause of lung cancer is that the RR for heavy smoking is ~25. It is extremely unlikely that such a large RR could be explained by some other confounding factor.

For randomized trials the interpretation of the RR is different – this is partly due to the fact that bias is substantially reduced by the RCT design itself, but also because, clinically, large treatment effects are just not very common. So for RCT studies, a general rule for an intervention that increases risk of an outcome is that RR values of ~1.10 are regarded as *small*, values of 1.2-2.0 are *moderate*, and those >2.0 are *large treatment* effects. For an intervention that decreases risk the equivalent RR estimates for small, medium and large effects are ~0.9, 0.5-0.8, and < 0.5, respectively. Again these same rules also apply to the OR.

ii. <u>The Relative Risk Reduction (RRR)</u>

The Relative Risk Reduction (RRR) is a measure of effect that is commonly used in the context of the RCT. The RRR is nothing more than a re-expression or re-scaling of the information provided by the RR. However, in clinical environments the RRR has more direct meaning than the RR since it indicates by how much in <u>relative</u> terms the event rate is decreased by treatment.

The RRR is calculated by as the 1 – RR or by dividing the *absolute risk reduction* [ARR] (see below) by the baseline risk (Risk$_{c)}$:

$$RRR = 1 - RR \quad \text{or} \quad \frac{ARR}{Risk_c} = \frac{Risk_c - Risk_t}{Risk_c}$$

In our RCT example, the RRR is therefore 1 – 0.62 = 38% (or 0.45 - 0.28 / 0.45 = 38%). That is, the death rate is 38% lower after ligation treatment, compared to sclerotherapy treatment. Thus the RRR represents the proportion of the original baseline (or control) risk that is removed by treatment.

The RRR is applied in the context of a treatment that reduces the risk of some adverse outcome. The RRR indicates the *magnitude of the treatment effect* in relative terms. As a general rule, treatments with a RRR of 10% or less are regarded as having a *small* effect, those with a RRR of 10-30% are *moderate*, and those >30% are regarded as *large* treatment effects. (Note that in absolute terms, the effect of treatment would depend on both the RRR and the baseline risk - as quantified by the absolute risk reduction)

iii. **The Absolute Risk Reduction (ARR)**

A more clinically useful measure of the effect of a treatment is the *absolute risk reduction* [ARR] (also confusingly referred to as the *attributable risk* in the Fletcher text and may be called the risk difference (RD) elsewhere). The ARR is simply the absolute difference in risks between the control and treatment groups.

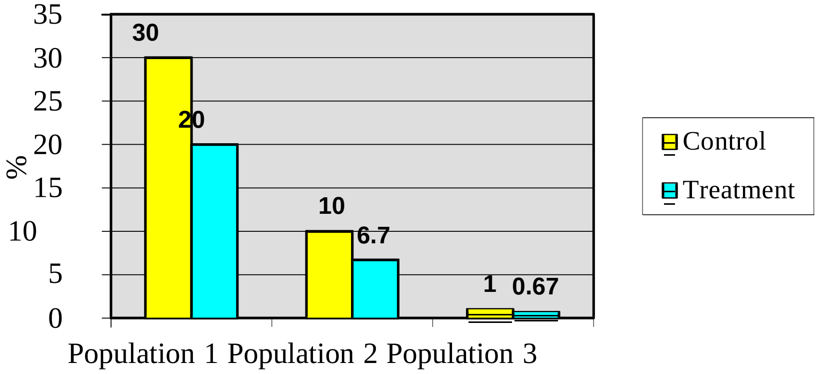
$$ARR = Risk_c - Risk_t$$

For the RCT example, the ARR is therefore 0.45 – 0.28 = 0.17 or 17%. The ARR represents the absolute difference in the risk of death between the two treatment groups. So the risk of death is 17% lower with ligation treatment, compared to sclerotherapy treatment. Like all difference measures the null value of the ARR is 0 (i.e., the point estimate that indicates no increase or decrease in risk).

The ARR is a simple and direct measure of the impact of treatment – in this example it tells you that 17% more patients treated with ligation will survive compared to using sclerotherapy. Contrast this information with that provided by the RR which tells you that ligation results in a death rate about 2/3^{rds} that of sclerotherapy – the RR does not tell you about the *absolute* benefit or effect of the treatment (only its *relative* benefit). It is for this reason that the ARR is the preferred measure when discussing the benefits of clinical interventions at the individual patient level.

A critically important point about the ARR is that it will vary depending on what the baseline risk is in the control group ($Risk_{c)}$, as illustrated in Figure 1 below. This figure shows the ARR for the same treatment effect (RRR = 0.33) for three populations that have baseline (control) risks or event rates of 30%, 10%, and 1%, respectively. Note that the ARR, which indicates the absolute impact of the treatment, varies from 10% in population 1 to a meager 0.33% on population 3. Thus the absolute benefit of treatment depends upon how much risk there is in the population before the treatment is applied (i.e., the baseline risk). When the baseline risk is high (as in population 1) the treatment can have a large
impact (it will reduce the number of subjects who have the outcome by 10%), whereas when the baseline risk is low (population 3) the effect of treatment is minimal (it reduces the number of subjects who have the outcome by only 0.33%).

Figure 1. Effect of different baseline risks on the ARR given a common RRR of 0.33



ARR= 10% **ARR= 3.3%** **ARR = 0.33%**

Since the magnitude of the baseline (control) event rate can vary widely from one population to another or from one age/sex/gender group to another, the ARR for a given treatment can vary markedly from one clinic or hospital to another. The bottom line is that an ARR calculated in one study or for one population **cannot** be directly applied (or transported) to another population.

It is important to note that in contrast to the ARR we make an explicit assumption that the relative impact of an intervention (as measured by the RRR) is regarded as a constant entity – i.e., we assume that it does not change from one population to another. Thus, in the case of the treatment of esophageal varices we assume that ligation treatment reduces the relative risk of death by 38% in all populations, whereas the absolute effect (as measures by the ARR) will be dependent on the baseline risk in the population. The assumption of constant RRR across all populations can of course be challenged as it is likely that the effect of treatment would differ across widely different populations.

One last point about the ARR is that it is used when the treatment group results in a lower event rate than the control group (i.e., when the RR is < 1.0), which is typical of therapeutic trials where the treatment is designed to reduce the occurrence of a bad outcome. However, treatments can also have harmful side effects. In this case we would see a RR of > 1, and would calculate an **ARI or Absolute Risk Increase** which would indicate in absolute terms how much more harmful events are seen in the treatment group. In turn, the ARI is used to calculate a NNH (Number Needed to Harm), similar to the concept of the number need to treat explained below. The potential confusion caused by referring to ARR or ARI is one of the reasons that the term risk difference (RD) – the absolute difference between two risks – is preferred by some epidemiologists.

iv. The Number Needed To Treat (NNT)

The *number need to treat* (NNT) illustrates the number of patients who would need to be treated in order to prevent one adverse event. It is calculated simply as the inverse of the ARR:

$$\text{NNT} = \quad 1 / \text{ARR}$$

Using our RCT example, the NNT is 1/ 0.17 = 5.9 (or 6). This means that for every 6 patients who received ligation treatment rather than sclerotherapy treatment, one death is prevented. The NNT can be seen as a simple re-expression of the information provided by the ARR, however, people have a hard time interpreting absolute probabilities (like an ARR of 17%). So converting these probabilities into "real numbers" (as the NNT does) provides more readily interpretable information. Note that since the ARR increases with increasing time it will have a concomitant effect on the NNT (which will get smaller), also just like the ARR, the NNT will be influenced by differences in baseline event rates. Thus to be interpretable the NNT should always be accompanied by a specific time period e.g., 1-year NNT = 25, 5-year NNT = 5 etc. (note that in our sclerotherapy RCT the mortality rates were estimated for the average duration of follow-up in the study which was 10 months. So the NNT of 6 should really be described as a 10-month NNT of 6 for survival).

The NNT is useful because it conveys the amount of work required to take advantage of the potential clinical benefit of an intervention. A high number indicates that a lot of effort will be expended to gain any benefit. For example, the 1-year NNT for primary stroke prevention in adults < 45 years of age using statin drugs is a staggering 13,000 – meaning that 13,000 patients would have to be treated with statins for one year to prevent one stroke event. However, the NNT for secondary stroke prevention (that is, among patients who have already suffered a stroke) using statins is only 57. Why would the NNT for the same drug be so different between these two applications?

The value of the NNT depends on the relative efficacy of the intervention (as indicated by the RRR) and the underlying baseline risk. This is demonstrated in the following table, notice the wide range of NNT values (Laupacis et al, NEJM 1988;318:1728-33):

Table. Effect of base-line risk and relative risk reduction on NNT

Base-line Risk (%)	**Relative Risk Reduction (RRR)**			
	0.5	**0.25**	**0.20**	**0.10**
60	3	7	8	17
30	7	13	17	33
10	20	40	50	100
5	40	80	100	200
1	200	400	500	1000
0.1	2000	4000	5000	10000

A good clinical example of the calculation and use of NNT and NNH measures to understand the risk and benefits of a treatment is provided by the CHARISMA trial (Bhatt DL et al. *N Engl J Med.* 2006;354:1706-1717). In this randomized, placebo controlled, multi-center clinical trial, 15,603 with cardiovascular disease or multiple risk factors were randomized to 2 different anti-platelet regimens: 1) Clopidogrel (75 mg/d) and low dose aspirin (ASA) (72-162 mg/d) or 2) low dose ASA and placebo. The composite outcome was any stroke, MI or CV death. The median follow-up period was 28 months. In the group that got ASA only, 7.3% of 7,801 subjects had an event during follow-up, while 6.8% of the 7,802 subjects randomized to the Clopidogrel and ASA group had an event (RR = 0.93, 95% CI 0.83-1.05, p = 0.22). The ARR was therefore 0.5% and so the NNT was 200. In other words 200 patients would need to be treated for 2.3 years with Clopidogrel and ASA to prevent one additional major outcome (compared to ASA alone). However, the Clopidogrel group had a higher rate of bleeding complications – 1.7% had severe or fatal bleeding compared to 1.3% of the ASA only group (RR = 1.25, 95% CI 0.97-1.61, p = 0.09). This results in an ARI of 0.4% and a NNH of 250. So for every 250 patients treated with Clopidogrel and ASA for 2.3 years, one extra severe or fatal bleed will occur (compared to ASA alone). The NNT of 200 and NNH of 250 can therefore help to describe in absolute terms the trade off in benefits and risks of adding Clopidogrel to low dose ASA for the prevention of CVD. Would you advise a family member to go onto Clopidogrel if this was recommended to them? Why or why not?

II. Population-based Measures of Effect (PAR, PARF)

Population Attributable Risk (PAR) and Population Attributable Risk Fraction (PARF)

In terms of understanding the impact of a risk factor on the incidence of disease in the population at large it is necessary to know both the relative effect of the risk factor on disease risk (i.e., the RR), as well as the prevalence of the risk factor in the population. Clearly, a risk factor would be expected to result in more disease in a population if it is both strongly associated with disease risk (i.e., has a large RR) and is more common within the population. Two measures, the *Population Attributable Risk* (PAR) and the *Population Attributable Risk Fraction* (PARF), are used to quantify the impact of a risk factor on disease at the population level under the implicit assumption is that the risk factor is a *cause* of the disease.

The *Population Attributable Risk* (PAR) represents the excess disease in a population that is associated with a risk factor. It is calculated from the absolute difference in disease risks between the exposed and non-exposed groups (i.e., the risk difference [RD]) and the prevalence (Prev) of the risk factor in the population i.e., PAR = RD x Prev. The PAR represents the excess disease (or incidence) in the population that is caused by the risk factor.

The *Population Attributable Risk Fraction* (PARF) represents the fraction of total disease in the population that is attributable to a risk factor. It is calculated as the PAR divided by the total incidence in the population. Thus it represents the proportion of the total incidence in the population that is attributable to the risk factor i.e., PARF = PAR/Total incidence. The PARF represents the maximum potential impact of prevention efforts on the incidence of disease in the population if the risk factor were eliminated. Again, the implicit assumption is that the risk factor is a *cause* of the disease, and that removing it would reduce the incidence of disease.

Example: Lung cancer mortality and smoking in British doctors (Data from earliest report on the cohort published by Doll and Hill, BMJ 1964 – See Table 5.4 in FF).
Total mortality rate from lung cancer= 0.56/1,000 person years
Mortality rate from lung cancer in ever smokers = 0.96/1,000 person years
Mortality rate from lung cancer in never-smokers = 0.07/1,000 persons years RR
of ever smoking and lung cancer death = 0.96/0.07 = 13.7
Prevalence of smoking (among British doctors in the 1950's) = 56%

PAR = RD x Prev = [0.96 – 0.07] x 0.56 = 0.50/1,000 person years (or 0.05% per year). This represents the absolute excess of lung cancer mortality among the doctors due to smoking.

PARF = PAR/Total mortality = 0.50/0.56 = 89%. This represents the proportion of all lung cancer deaths in this population that is due to smoking (it indicates that the vast majority of lung cancer death is caused by smoking alone).

The PARF can also be calculated directly if the RR and prevalence are known using the following equation:

$$\text{PARF} = \frac{\text{Prev.(RR-1)}}{1 + \text{Prev.(RR-1)}}$$

So, if the RR= 13.7 and Prev = 56%, then the PARF = [0.56(13.7-1)]/[1 + 0.56(13.7-1)] = 88% (slight difference is due to rounding). Note that the potential impact of prevention efforts can be gauged by calculating the PARF using lower estimates of prevalence. For example if the prevalence of smoking was reduced to, say, only 10%, the PARF of smoking for lung cancer mortality would be reduced to 56% i.e., PARF = [0.10(13.7-1)]/[1 + 0.10(13.7-1)] = 55.9%.

The PAR and PARF indicate the potential public health significance of a risk factor. For example, a large risk that is very rare is unlikely to cause much disease in the population (and therefore has less public health significance), as opposed to a small risk that is very common which would lead to a high PARF with more public health significance. For example, a risk factor that has a big effect (i.e., RR= 10) but is rare (P = 0.1% or 0.001) has a PARF of 1%, whereas a risk factor that has a small effect (i.e., RR = 2) but is common (P = 40% or 0.4) has a PARF of 44%. (See the lecture on Cohort Studies for further discussion of the PARF).

III. The Odds Ratio (OR)

The Odds Ratio (OR) is the measure of effect of choice for case control studies (CCS), because the CCS design is not able to quantify the actual incidence or risk of disease in exposed and non-exposed groups. (see CCS lecture for further discussion) The OR is usually a good approximation of the RR – and it represents a clever fix to get around the problem that the case control design cannot generate the RR. As the name implies the OR is a ratio of odds, specifically, the odds of exposure in cases compared to the odds of exposure in controls. Like the RR the OR describes the *magnitude or strength* of an association between an exposure and the outcome of interest and like the RR its null value is 1.0. An OR > 1.0 indicates a positive association between the exposure and disease, while an OR < 1.0 indicates a negative association.

As an example of the calculation of an OR, imagine we had done a case control study rather than a cohort study to assess the relationship between ever smoking and death from lung cancer in the British doctors. We assess the smoking status in 100 lung cancer deaths (the cases) and 100 doctors who were alive or had died of something other than lung cancer (the controls). We get the following data:

Example – CCS to measure the association between smoking and lung cancer death in British doctors

	Outcome	
	Lung CA Death	No Lung CA Death
Smoker	95 (a)	56 (b)
Non-smoker	5 (c)	44 (d)
	Odds of exposure 95/5	Odds of exposure 56/44

The OR is calculated as the ratio of the odds of exposure in the cases (i.e., 95/5) divided by the odds of exposure among the controls (i.e., 56/44) which equals 15.6. In this example the OR is a very good approximation of the RR that was generated from the cohort study design (i.e., 13.7). One would interpret the OR as saying that the odds of death due to lung cancer was 15.6 times higher in smokers compared to non-smokers. As a general rule the OR more closely approximates the RR when the outcome of interest is rare in the underlying population (i.e., <10%). In this case, the rate of lung cancer mortality in the overall population is very rare i.e., 0.56 per 1,000 persons/year (equivalent to 0.056% per year). The fact that the OR is a good estimate of the RR when the event or disease is rare also makes sense given the data shown in Table 1 in the Frequency Lecture notes, which shows that the odds and probability are more alike when the risk is small (<10%).

Note that the OR was calculated using the numbers in the 4 cells labeled a, b , c, d – specifically, a/c divided by b/d. The equation simplified to (a x d)/(b x c) which is known as the cross-product ratio. One advantage of the OR is that it is <u>symmetrical,</u> meaning that rather than calculating the odds of exposure in the cases relative to the controls, if you compare the odds of disease among the exposed (a/b) relative to the odds of disease amongst the non-exposed (c/d) you end up with the same OR (since both approaches reduce to the ad/bc cross product ratio). One should note that the RR does not have this same characteristic of symmetry and that flipping the definition of outcome (from, for example, 'bad' to 'good") can often result in completely different results being obtained from the RR. For example, a significant RR for a 'bad' outcome such as death (e.g., RR= 0.60, 95% CI 0.45-0.75) can be converted to a non-significant result for the complementary 'good' outcome of survival (e.g., RR= 1.32, 95% CI 0.88-2.10) – even though the data itself has not changed!

While the OR is the effect measure of choice in CCS designs (See lecture on CCS design for more details), another potential advantage of the OR is that it can be used for any other type of study design i.e., cross-sectional studies, cohort studies, RCTs, and meta-analyses. Although the OR is widely used across all sorts of designs the savvy reader should not forget that the OR has several disadvantages over the RR. These include:

1) For cohort studies and trials the OR deviates from the true RR as the baseline risk in the untreated group (CER) increases – the effect is noticeable once the risk is > 10% which is often the case in RCTs, and the deviation can be substantial when baseline risk gets to be > 50%. This deviation is driven in part by the mathematical fact that the upper limit RR is limited by the baseline or control event rate (i.e., max RR = 1/CER). So, for example, if the CER is 50% the maximum possible value for the RR is 2.0.

2) The OR deviates from the true RR as the treatment effect gets larger – the effect is particularly noticeable once the RR reaches < 0.75 or greater (or equivalently >1.33).

3) The OR is <u>always</u> further away from the null value (1.0) than the RR – thus the treatment effect is always over-estimated by the OR compared to the RR.

4) The odds ratio can only be interpreted like a RR when it is a good approximation of the RR. So when the baseline risk is < 10% it is acceptable to talk about an OR as if it were a RR. So for example if the OR was 1.5 it would be permissible to describe this in terms of "*the likelihood or risk of disease was 50% higher in the exposed group*". However, when the OR is not a good approximation to the RR (e.g., the baseline risk is > 10%) then the OR should be described as an OR (which it is!) and <u>not</u> as if it were a RR! So in this example we would describe the OR of 1.5 as, "*the <u>odds</u> of disease was 50% higher in the exposed group*" (note we do not say the risk or likelihood).

5) As might be gleaned from the above discussion there is nothing clinically intuitive about the OR!

These disadvantages imply that when the data can be specified in terms of the RR i.e., the RCT or cohort design, then the RR should be the effect measure of choice.

EPI-546 2016

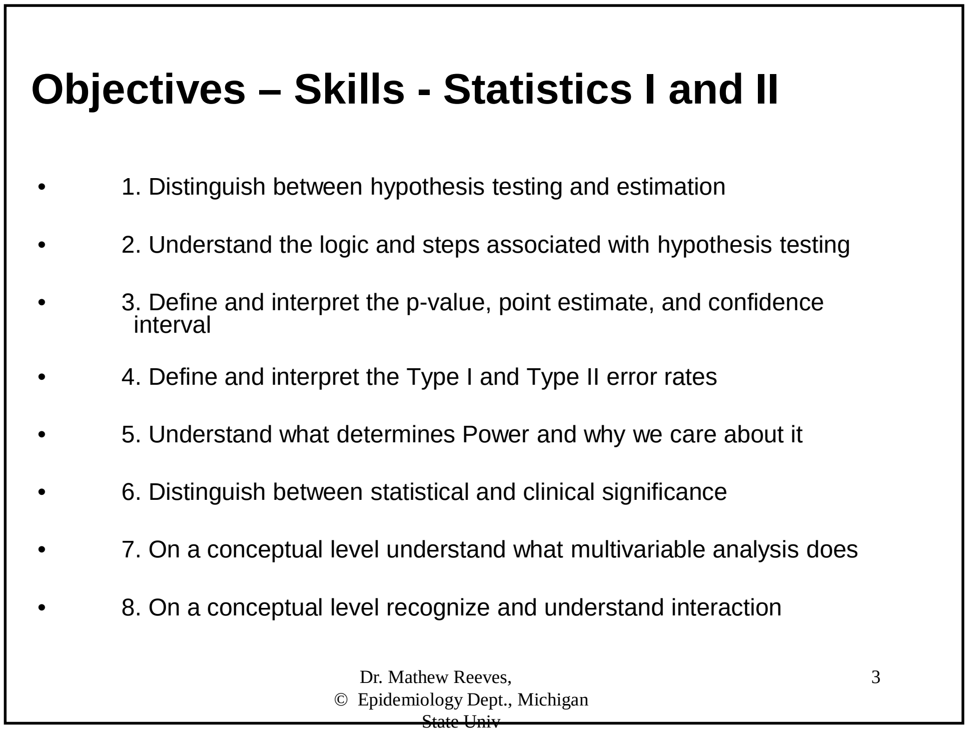
Lecture - Statistics I
Hypothesis testing

- Is there a significant difference?

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology

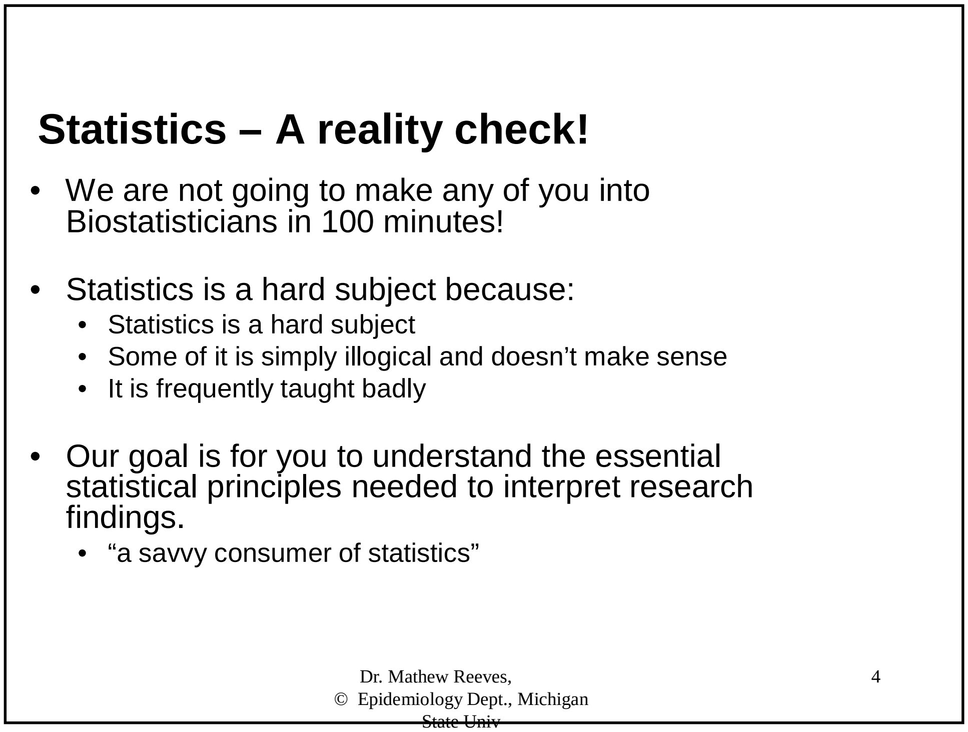
Objectives – Concepts- Statistics I and II

- 1. Concept of sampling
- 2. Systematic vs. random error
- 3. Two approaches to statistical inference
  - Hypothesis testing vs. estimation
- 4. Hypothesis (significance) testing
  - Null vs. alternative hypothesis
  - P-values and statistical significance
  - Type I (alpha) and Type II (beta) error rates
  - Power and sample size estimation
- 5. Estimation
  - Limitations of the p-value
  - Point estimates and Confidence Intervals
- 6. Clinical vs. statistical significance
- 7. Multiple comparisons
- 8. Multivariable analysis and interaction



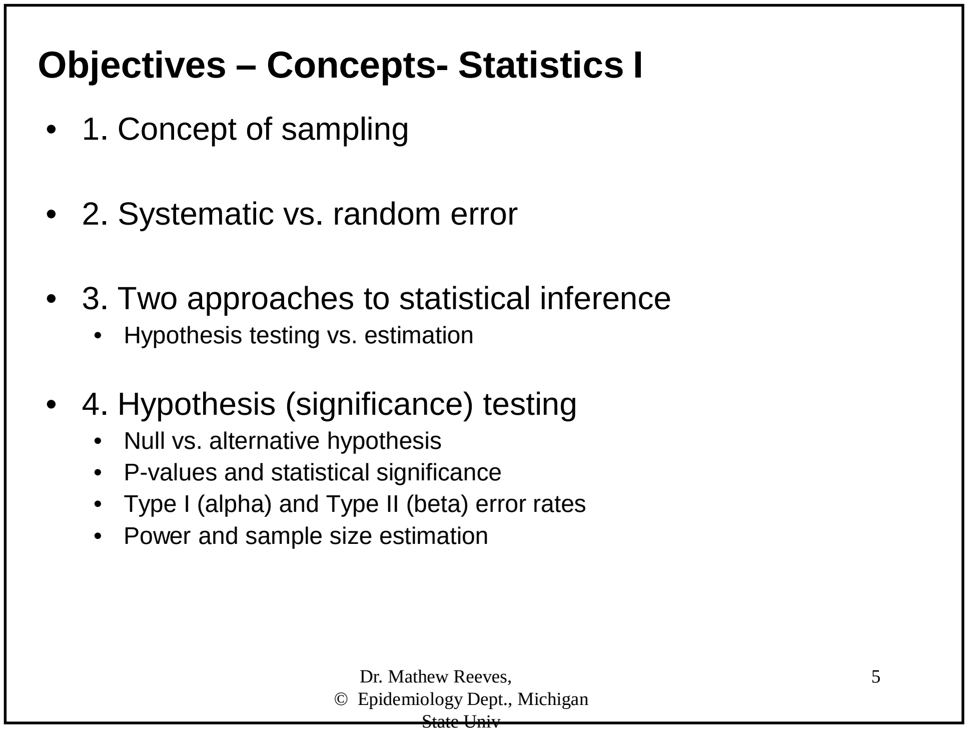
Objectives – Skills - Statistics I and II

- 1. Distinguish between hypothesis testing and estimation
- 2. Understand the logic and steps associated with hypothesis testing
- 3. Define and interpret the p-value, point estimate, and confidence interval
- 4. Define and interpret the Type I and Type II error rates
- 5. Understand what determines Power and why we care about it
- 6. Distinguish between statistical and clinical significance
- 7. On a conceptual level understand what multivariable analysis does
- 8. On a conceptual level recognize and understand interaction



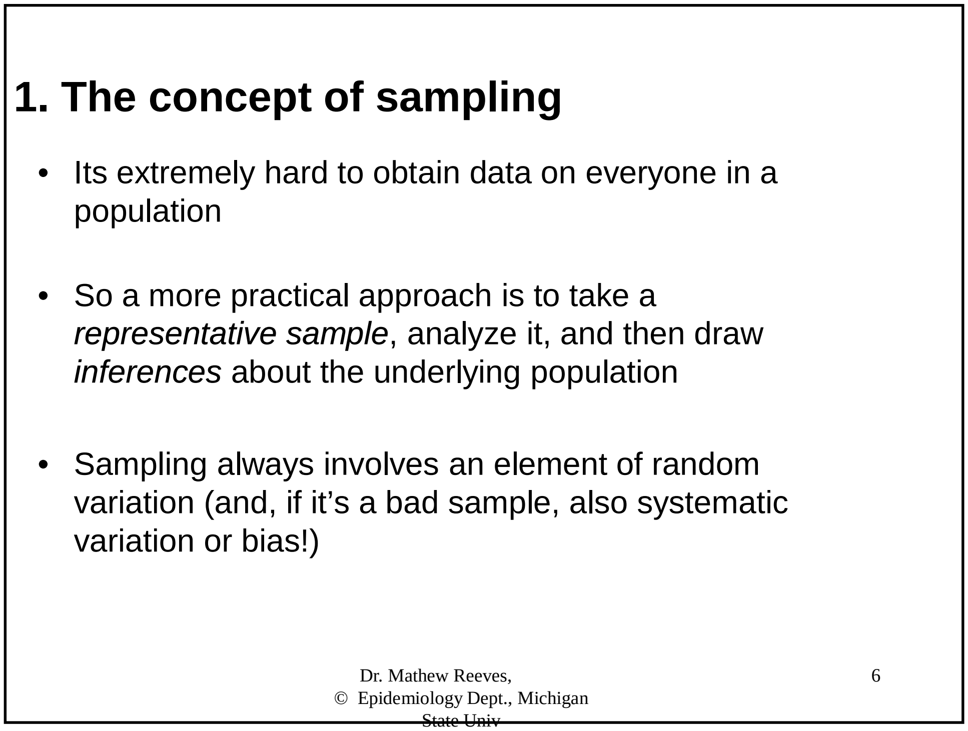
Statistics – A reality check!

- We are not going to make any of you into Biostatisticians in 100 minutes!
- Statistics is a hard subject because:
  - Statistics is a hard subject
  - Some of it is simply illogical and doesn't make sense
  - It is frequently taught badly
- Our goal is for you to understand the essential statistical principles needed to interpret research findings.
  - "a savvy consumer of statistics"



Objectives – Concepts- Statistics I

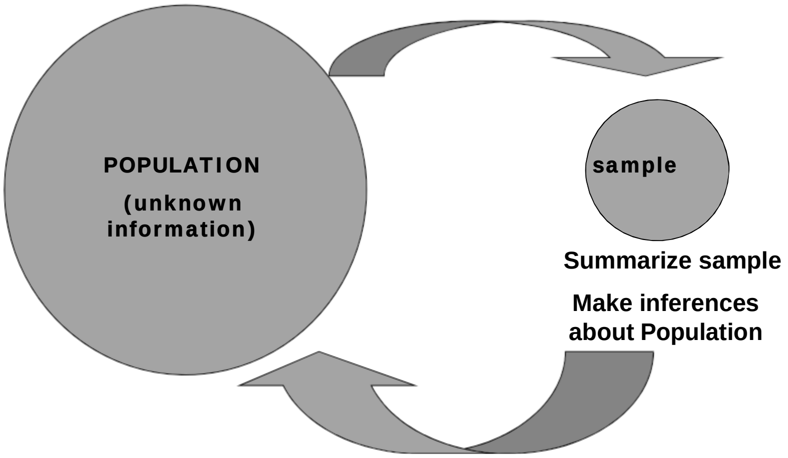
- 1. Concept of sampling
- 2. Systematic vs. random error
- 3. Two approaches to statistical inference
  - Hypothesis testing vs. estimation
- 4. Hypothesis (significance) testing
  - Null vs. alternative hypothesis
  - P-values and statistical significance
  - Type I (alpha) and Type II (beta) error rates
  - Power and sample size estimation



1. The concept of sampling

- Its extremely hard to obtain data on everyone in a population
- So a more practical approach is to take a *representative sample*, analyze it, and then draw *inferences* about the underlying population
- Sampling always involves an element of random variation (and, if it’s a bad sample, also systematic variation or bias!)

Sampling



POPULATION

(unknown information)

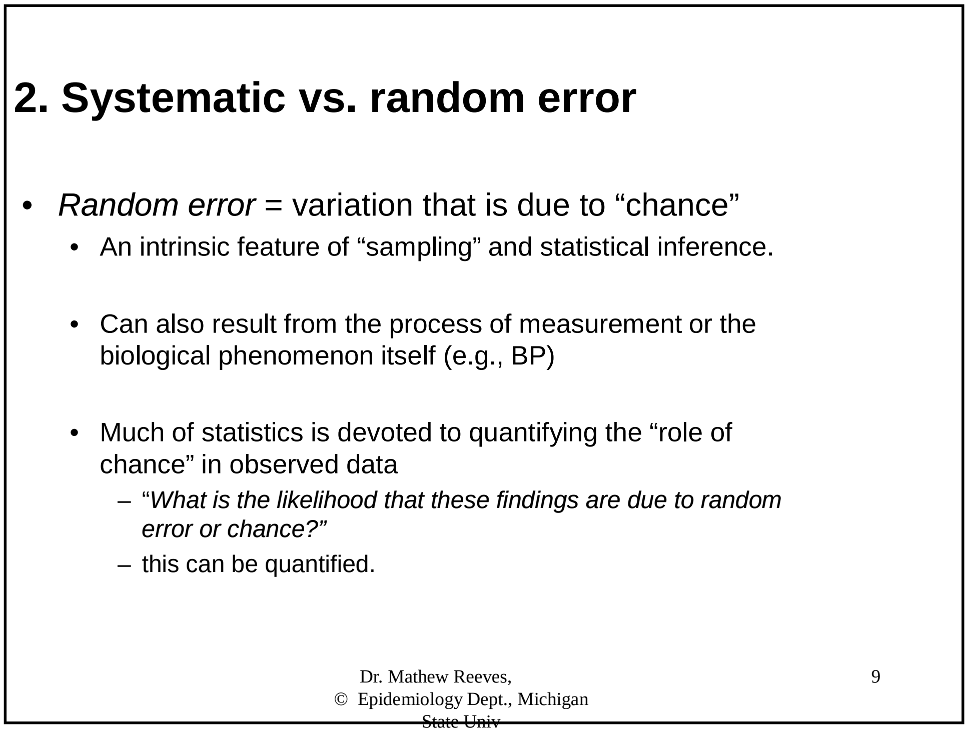
sample

Summarize sample

Make inferences about Population

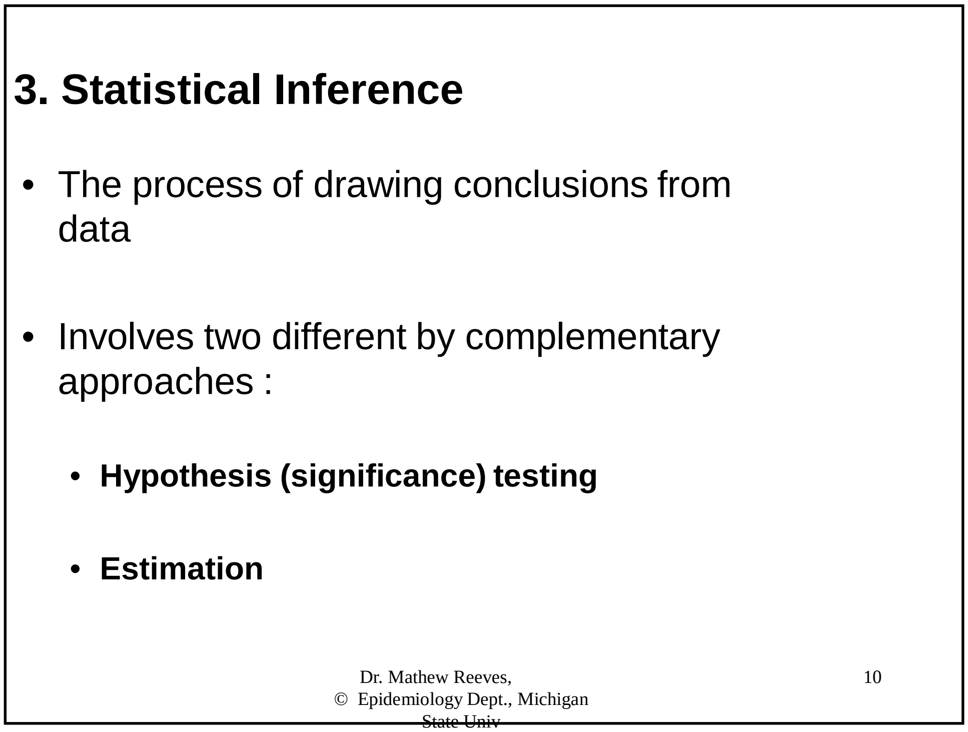
2. Systematic vs. random error

- Systematic error = Bias
  - *Defn: Any process that acts to distort data or findings from their true value*.
  - a.k.a. *validity* or *accuracy*
    - both terms imply a lack of systematic error
  - Hard to quantify in absolute terms (hence not the major focus of statistics) but if often far more important than random error.
  - Categorized into selection bias, measurement bias or confounding bias
    - See later lectures and EPI-547



2. Systematic vs. random error

- *Random error* = variation that is due to “chance”
  - An intrinsic feature of “sampling” and statistical inference.
  - Can also result from the process of measurement or the biological phenomenon itself (e.g., BP)
  - Much of statistics is devoted to quantifying the “role of chance” in observed data
    - “*What is the likelihood that these findings are due to random error or chance?”*
    - this can be quantified.



3. Statistical Inference

- The process of drawing conclusions from data
- Involves two different by complementary approaches :
  - **Hypothesis (significance) testing**
  - **Estimation**

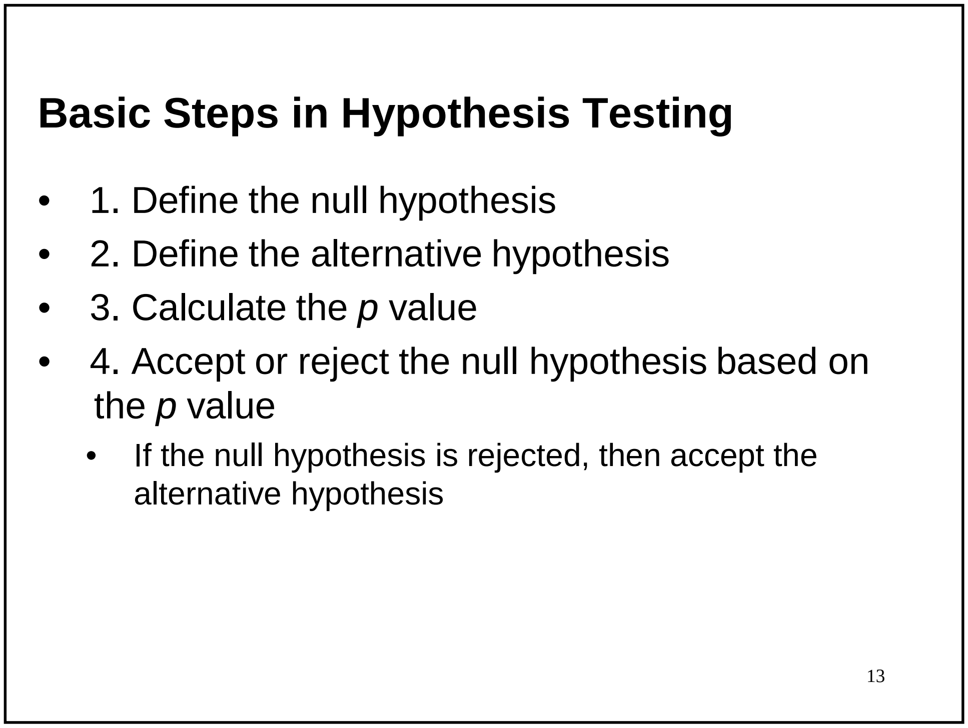
Hypothesis testing vs. Estimation

- **Hypothesis (significance) testing**
  - Concerned with making a *decision* about a hypothesized value of an unknown parameter
  - Involves the use of the p-value.
  - Views experimentation as **decision making**
  - *"Should I prescribe drug A or drug B?"*

- **Estimation**
  - Concerned with estimating the *specific value* of a unknown parameter
  - Involves the use of the confidence interval (CI)
  - Views experimentation as a measurement exercise
  - *"What did you find and how precisely did you measure it?"*

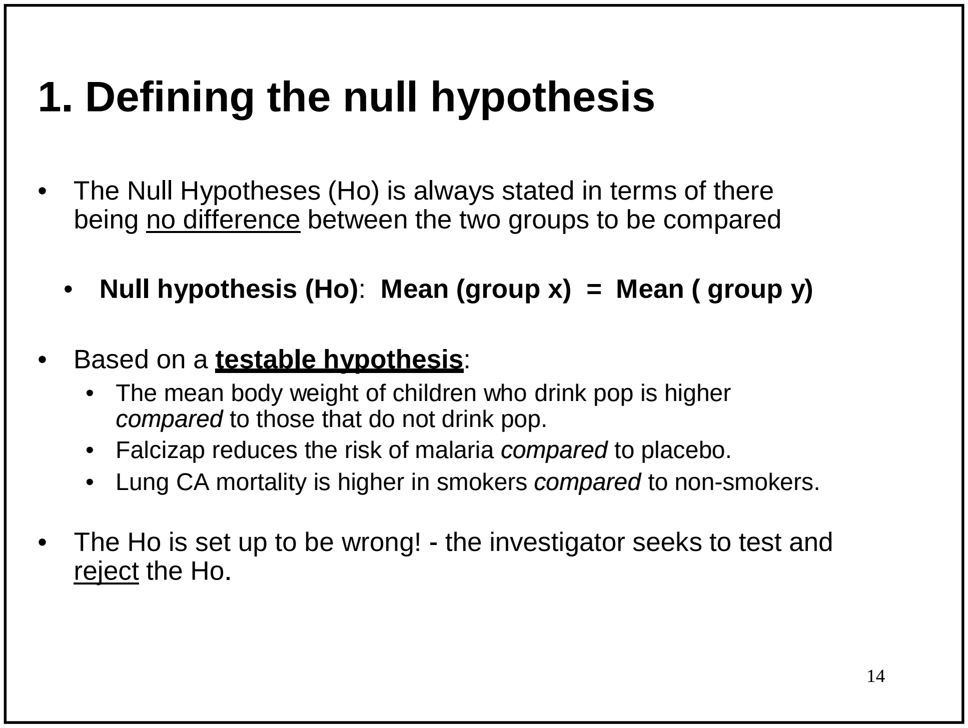
Sir Ronald Aylmer Fisher (1890- 1962)

- Sir Ron was an English statistician, evolutional biologist, eugenicist, and geneticist.
- Graduated from Cambridge in 1912
- Worked at Rothamsted Agricultural Experimental Station. Invented ANOVA.
- In 1935 published *The Design of Experiments.*
  - "*a genius who almost single- handedly created the foundations for modern statistical science*"
- Did not believe early data on smoking and lung cancer!



Basic Steps in Hypothesis Testing

- 1. Define the null hypothesis
- 2. Define the alternative hypothesis
- 3. Calculate the *p* value
- 4. Accept or reject the null hypothesis based on the *p* value
  - If the null hypothesis is rejected, then accept the alternative hypothesis



1. Defining the null hypothesis

- The Null Hypotheses (Ho) is always stated in terms of there being <u>no difference</u> between the two groups to be compared
  - **Null hypothesis (Ho)**: **Mean (group x) = Mean (group y)**
- Based on a **<u>testable hypothesis</u>**:
  - The mean body weight of children who drink pop is higher *compared* to those that do not drink pop.
  - Falcizap reduces the risk of malaria *compared* to placebo.
  - Lung CA mortality is higher in smokers *compared* to non-smokers.
- The Ho is set up to be wrong! - the investigator seeks to test and <u>reject</u> the Ho.

2. Defining the alternative hypothesis

- The alternative hypotheses (Ha) is stated in terms of there being a difference between the two groups being compared
- **Alternative hypothesis (Ha)**: **Mean (group x)** ≠ **Mean (group y)**
- The **only way** that the Ha can be accepted is if we reject the Ho.
- We can never prove the Ha is true - we can only say that the Ho is false!

15

Example: Null and Alternative Hypotheses

- Testable Hypothesis
  - Lung CA mortality is higher in smokers *compared* to non-smokers
- Null hypothesis (Ho) = there is no difference in lung CA mortality between smokers and non-smokers
- If we reject the null hypothesis then we believe the alternative hypothesis
- Alternative hypothesis (Ha) = there is a difference in lung CA mortality between smokers and non-smokers

16

Isn’t it a bit odd that we have to test the Ho to prove the Ha?

- Sort of…..
- But the only scientific process we know to make valid conclusions is to test an already existing hypothesis (or decision)
- Example: NFL
  - *The Steelers are* ***challenging the ruling on the field*** *that the ball was caught*
  - Question: Is there enough video evidence to reject the ruling on the field (the Ho)?

3. Calculate the p-value
4. Accept or reject the null hypothesis

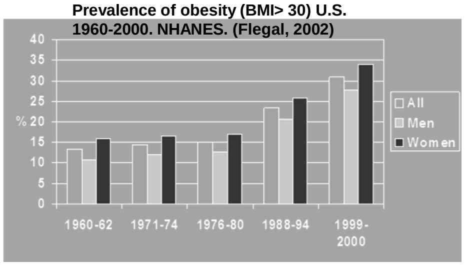
- *p* value= probability of obtaining the results observed, **if the null hypothesis were true**.

  - If *p* = 0.01, then the chance of obtaining the results if there was no difference between the groups is 1%
    - Thus the results are very unlikely to be due to chance
  - So we reject the null hypothesis and accept the alternative and conclude that the groups are different

- If *p* = 0.7, then the chance of obtaining the results if there was no difference between the groups is 70%
  - So the results are very consistent with the null hypothesis being true, and so we accept the Ho and conclude the groups are not different.

An example hypothesis test – the problem of Obesity

- Obesity is a big problem. The rates of obesity have more than doubled in the last 30 years
- Some people believe that one of the factors contributing to this is the large amounts of pop consumed - especially children .
- Our testable hypothesis is that children who drink pop are heavier than children who do not.
- We set out to test this hypothesis………….



Prevalence of obesity (BMI> 30) U.S. 1960-2000. NHANES. (Flegal, 2002)

19

An example hypothesis test

- **Null hypothesis (Ho)**: **Ux = Uy**
  - the mean body wt. of pop-drinking children (Ux**)** is <u>not</u> different from the mean body wt. of children who abstain (Uy**)**
- **Alternative hypothesis (Ha)**: **Ux ≠ Uy**
  - the mean body wt. of pop-drinking children is not equal to the mean body wt. of children who abstain
- <u>Let's set up an experiment:</u>
  - 40 children randomized to pop-drinking (the intervention group [X]) or no- pop drinking (the control group [Y])
  - Measure body weight at day 90
  - Calculate the mean difference in the body weights between group X and group Y
  - Determine whether a statistical difference exists between mean wt. of group X and group Y using an appropriate statistical test (in this case the **t-test)**

20

The t-test……

$$t = \frac{\overline{X} - \overline{Y}}{S\,\overline{x} - \overline{y}}$$

X = mean weight of the pop group
Y = mean weight of the control group

Where:

$$S\,\overline{x} - \overline{y} = \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right).\ S^2}$$

= standard error of the difference between two means.

S^2 = estimate of pooled population variance

- Larger values of 't' result in smaller p values which are more consistent with Ho being false

All else equal, the value of 't' increases with a bigger difference between the group means (numerator), or a smaller standard error (denominator)

- Smaller standard error come from larger n (bigger studies have higher power)

Determining "statistical significance"

- A small p value, for example p = 0.01, indicates one of two things - either:
  - i) a rare event has occurred

  OR

  - ii) The Ho is false

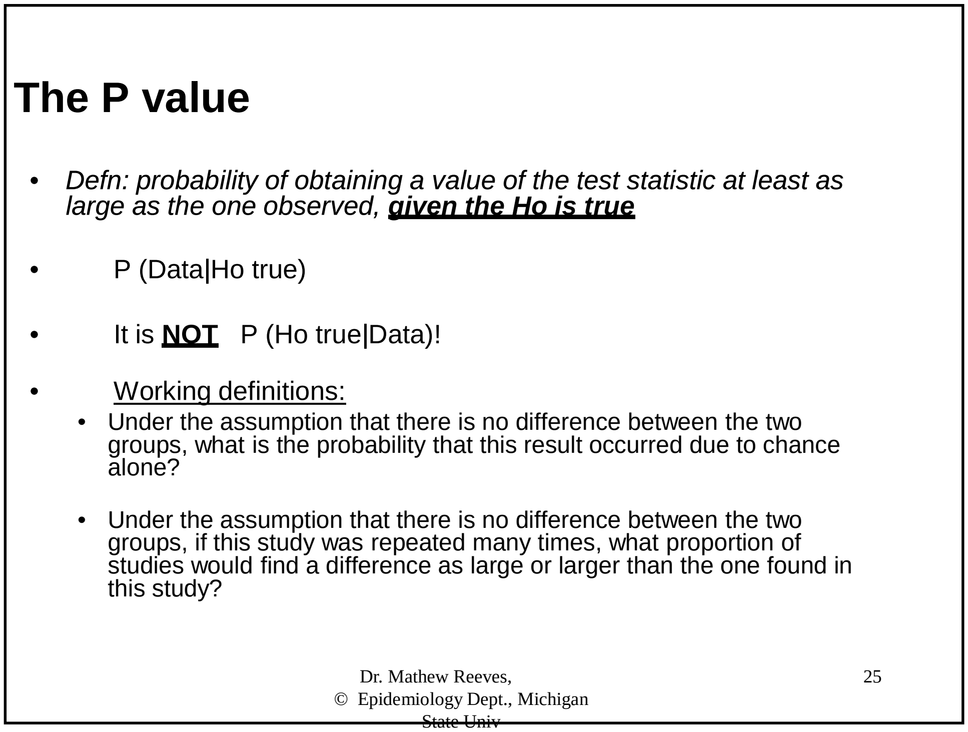
- By convention, the probability where we decide to reject the Ho is set at 0.05 or 5%. This is called the **significance level (or alpha)**

After 90 days of our pop drinking experiment…………

- The mean increase in body weight in group X (intervention) was 3.0 kg
- The mean increase in body weight in group Y (control) was 0.5 kg
- For this 2.5 kg difference the t-test calculated a p value of 0.02.
- *Practical Interpretation*
  - *Under the assumption that there is no difference between the two groups, the probability of observing an increase in body weight of >=2.5 kg by chance alone is only 2% (i.e., we would expect to see result as large or larger than this only two times out of 100 if there really was no difference)*

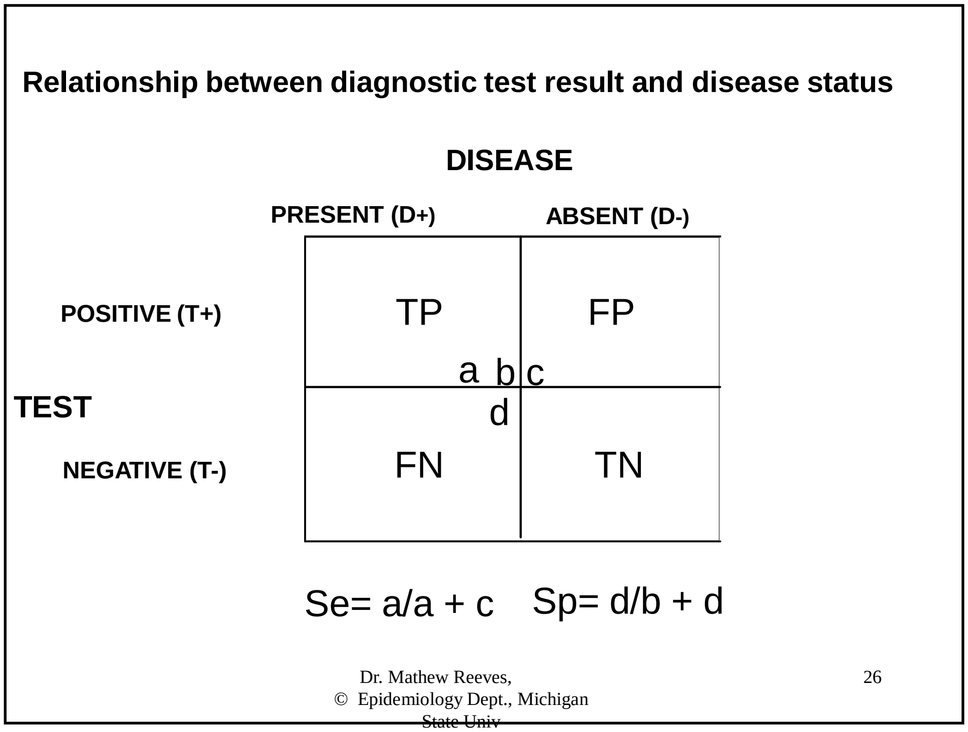
What do we conclude?

- Because 0.02 is smaller than 0.05 (the pre-defined significance level) we conclude the result is "statistically significant" and we therefore reject the Ho and accept the Ha
- By accepting the Ha we conclude that in this experiment the mean body weight of children randomized to pop-drinking was not equal to the mean body weight of children who were randomized to not drink pop
- Because this comparison was based on an experimental design (the RCT), bias is unlikely to be present (assuming we did a good job conducting the study) and so we can be confident that the exposure to pop caused the increase in body weight.



The P value

- *Defn: probability of obtaining a value of the test statistic at least as large as the one observed,* ***given the Ho is true***
- P (Data|Ho true)
- It is **NOT** P (Ho true|Data)!
- Working definitions:
  - Under the assumption that there is no difference between the two groups, what is the probability that this result occurred due to chance alone?
  - Under the assumption that there is no difference between the two groups, if this study was repeated many times, what proportion of studies would find a difference as large or larger than the one found in this study?



Relationship between diagnostic test result and disease status

		DISEASE	
		PRESENT (D+)	ABSENT (D-)
TEST	POSITIVE (T+)	TP	FP
	NEGATIVE (T-)	FN	TN

a b c
d

Se= a/a + c Sp= d/b + d

Type I and Type II errors

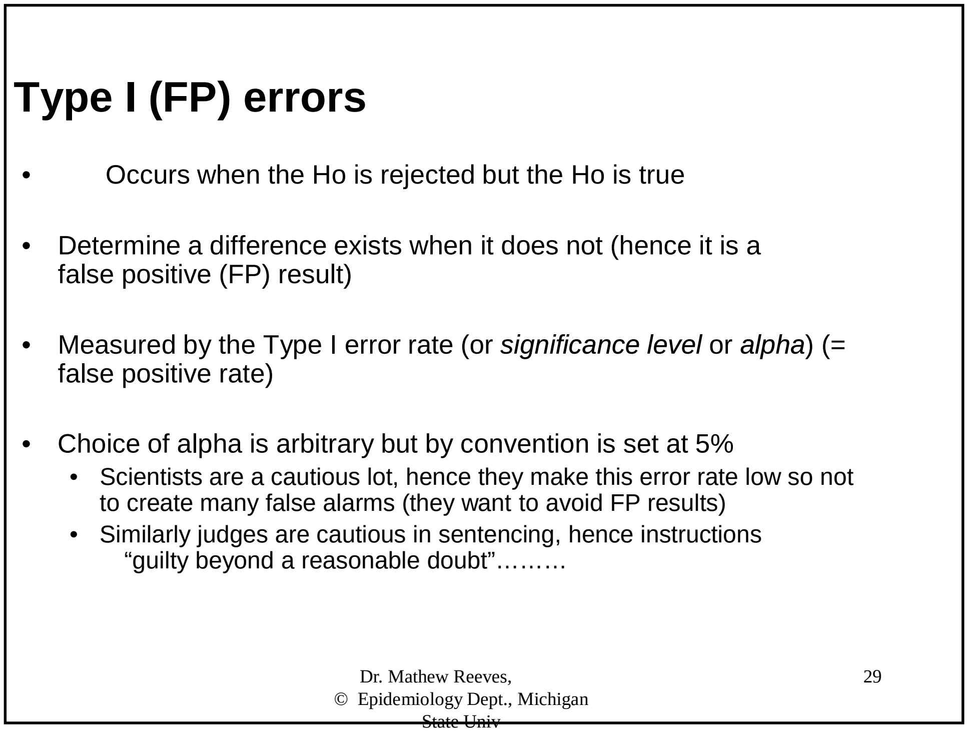
- Just as in diagnostic testing, statistical testing can result in errors.
- There are two types of errors one can make in statistical testing:
  - FP or Type I error
  - FN or Type II error

Relationship between significance test results and the truth

TRUTH

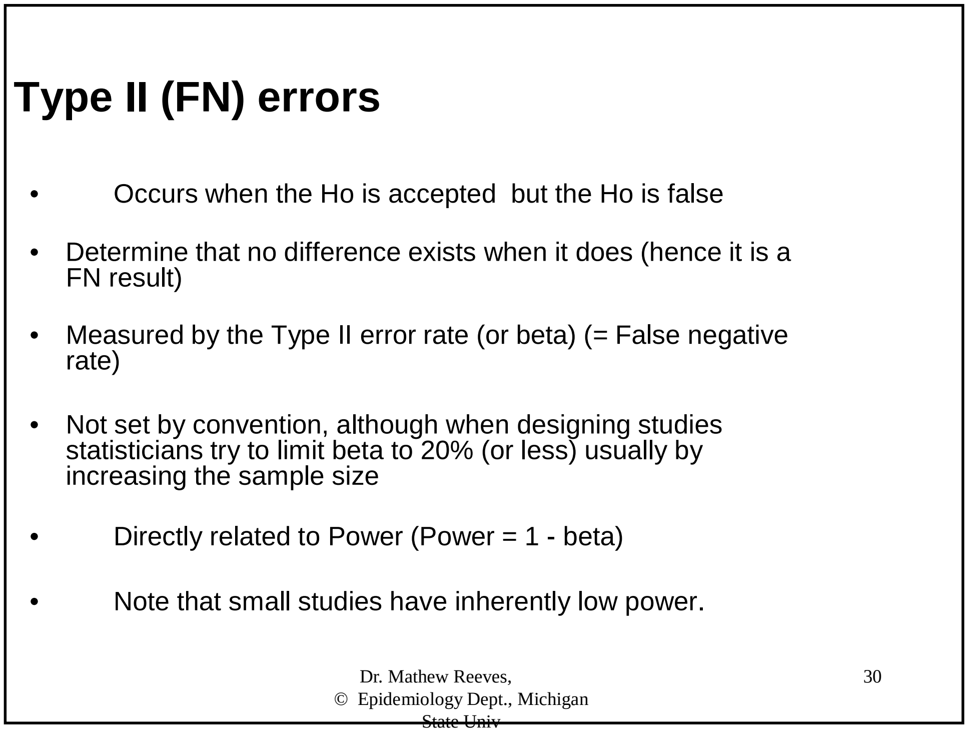
SIGNF TEST	**Ho False**	**Ho True**
REJECT Ho (P ≤ 0.05)	TP	FP Type I (a)
ACCEPT Ho (P > 0.05)	FN Type II (B)	TN

Power = (1 - B)



Type I (FP) errors

- Occurs when the Ho is rejected but the Ho is true
- Determine a difference exists when it does not (hence it is a false positive (FP) result)
- Measured by the Type I error rate (or *significance level* or *alpha*) (= false positive rate)
- Choice of alpha is arbitrary but by convention is set at 5%
  - Scientists are a cautious lot, hence they make this error rate low so not to create many false alarms (they want to avoid FP results)
  - Similarly judges are cautious in sentencing, hence instructions "guilty beyond a reasonable doubt"………



Type II (FN) errors

- Occurs when the Ho is accepted but the Ho is false
- Determine that no difference exists when it does (hence it is a FN result)
- Measured by the Type II error rate (or beta) (= False negative rate)
- Not set by convention, although when designing studies statisticians try to limit beta to 20% (or less) usually by increasing the sample size
- Directly related to Power (Power = 1 - beta)
- Note that small studies have inherently low power.

Power (1 - B)

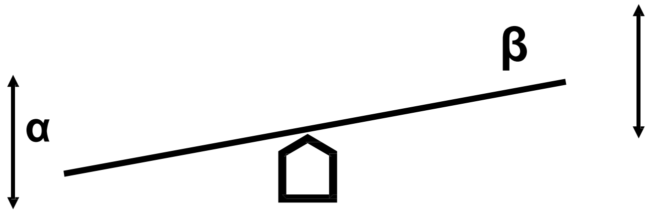
- *Defn: Probability of correctly rejecting Ho when Ho is false*
- Power is used in the planning phase of a study
  - try to have power of at least 80% (i.e., limit beta to 20% or less)
- If a study is designed to have 80% power, then it has an 80% chance of finding a significant difference if a difference exists
  - (i.e., an 80% chance of rejecting the Ho when the Ho is false)
- Power is analogous to the sensitivity of a diagnostic test

Power (1 - B)

- <u>Power</u> is a function of:
  - Alpha (FP) error rate
  - Beta (FN) error rate
  - Effect size
  - The Variability in the data

N.B. Alpha and beta are inversely related

(analogous to the trade off between Se and Sp)



<u>Power</u> is a function of:

- <u>Alpha (FP) error rate</u> (usually 5%)
  - Smaller the alpha error the harder it will be to identify a difference (because beta goes up and so power is *lower*)
  - So, if you wanted to be extra cautious and use alpha = 0.01, power would automatically be lower
- <u>Beta (FN) error rate</u>
  - Smaller the beta error the easier it will be to identify a difference (hence the *higher* the Power)
  - Power is usually manipulated by increasing the size of the study (more observations, bigger N) or by increasing alpha from 0.05 to 0.1 (or by picking a one-sided Ha)
  - Typically set at 20%, four times larger than alpha (this reflects the greater attention placed on FP vs. FN errors)

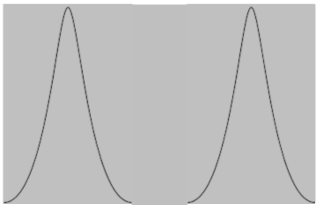
Power is a function of:

- Effect size
  - The magnitude of the difference you are trying to detect
  - Bigger differences are easier to detect (in statistics SIZE MATTERS!!)
  - When designing a study this is defined as the **minimal clinically important difference**
    - What is the smallest difference that would be important to know clinically?
    - e.g., In designing a BP reduction study what is a meaningful reduction in BP? Is a 0.5 mm Hg decline important? Or should it be 5 mm Hg?

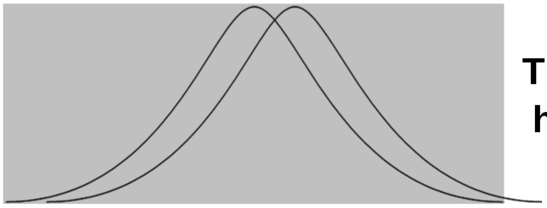
Power is a function of:

- The variability in the data
  - Continuous data
    - The greater the variability (a.k.a. dispersion or SD) in the data the harder it is to detect a difference
    - The more “noise” there is the harder it is to see the real “signal”
  - Categorical data (rates and proportions)
    - The rarer an event (e.g., death, relapse) the harder it is to detect a difference
    - The absolute number of events counted is more important than the underlying size of the study groups

The bigger difference and/or the less the variation the greater the Power



This difference is easy to detect (P < 0.05)



This is much harder

Consequences of Low Power Studies

- Difficult to interpret negative results:
  - truly no difference?, or did study simply fail to detect the true difference that exists?
- Failure to identify potentially important associations
- Low power means low precision
  - (as indicated by a wide confidence interval)

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes – Statistics I

Mathew J Reeves BVSc, PhD
Michael Brown MD, MSc

Objectives:

- 1. Distinguish between hypothesis testing and estimation
- 2. Understand the logic and steps associated with hypothesis testing
- 3. Define and interpret the p-value, point estimate, and confidence interval
- 4. Define and interpret the Type I and Type II error rates
- 5. Understand what determines Power and why we care about it
- 6. Distinguish between statistical and clinical significance
- 7. On a conceptual level understand what multivariable analysis does
- 8. On a conceptual level recognize and understand interaction

Outline:

I. Classical Hypothesis (significance) testing
  A. Ho, Ha, p value, alpha
  B. Summary - Steps in Classical Hypothesis Testing
  C. Common Statistical Tests
  D. Type I (alpha) error and Type II (beta) error
  E. Power and Sample Size
  F. Summary of terms and definitions

II. Estimation, Point Estimates and Confidence Intervals

III. Multiple comparisons

IV. Multivariable analysis and interaction.

Introduction

One of the most important tools available for improving the effectiveness of medical care is the information obtained from well-designed and properly analyzed clinical research. The purpose of this lecture is to teach concepts underlying appropriate statistical design and analysis of clinical and epidemiological studies, with an emphasis on using the clinical trial design as an example.

Statistical Inference is defined as the process of drawing conclusions from data. There are two different but complementary categories of statistical inference: hypothesis testing and estimation.

I. Classical Hypothesis (Significance) Testing

Hypothesis or significance testing is essentially concerned with making a *decision* about the value of an unknown parameter. It therefore views "experimentation" as a decision making exercise.

A. H_0, H_A, p value, alpha

Data from clinical trials are usually analyzed using *p* values and classical hypothesis testing. In classical hypothesis testing, two hypotheses which might be supported by the data are considered. The first, called the *null hypothesis* ($H_{0)}$, is the hypothesis that there is no difference between the groups being compared with respect to the measured quantity of interest. For example, in a study examining the use of a new sympathomimetic agent for blood pressure support in patients with sepsis, the null hypothesis is that there is no difference between the systolic blood pressure achieved with the new agent and the control agent. The *alternative hypothesis* (H_A) is that the groups being compared are different i.e., the blood pressure for the new agent is either higher or lower.

Null hypothesis (H_0): Ux = Uy

i.e., the mean blood pressure is NOT different between the new agent (x) and the control agent (y).

Alternative hypothesis (H_A): Ux $\neq$ Uy

i.e., the mean blood pressure is different between the new agent (x) and the control agent (y).

Sometimes the alternative hypothesis is specified in terms of the direction of the difference e.g., the blood pressure for the new agent is higher that the control agent (this is referred to as a one-sided alternative, as opposed to the two-sided alternative shown above). Regardless of whether a one-side or two sided alternative is used, the difference between the two groups is called the *treatment effect*.

Once the null and alternative hypotheses are defined, the null hypothesis is "tested" to determine whether it will be accepted as true, or whether it will be rejected and the alternative hypothesis accepted as true. The process of testing the null hypothesis consists of calculating the probability of obtaining the results observed (or results that are even more extreme), **assuming the null hypothesis is true**. This probability is the *p value*. The p value is defined as *the probability of observing the test statistic at least as large as the one observed under the assumption that the null hypothesis is true*. In terms of conditional probabilities we can write it as: P (Data|H_0 true).

If the *p* value is less than a predefined value, denoted as the *significance level* or *alpha* (α), then the null hypothesis is rejected, and the alternative hypothesis is accepted as true. By convention, in most clinical studies the significance level is set to 5%, but it can be changed according to how stringent (e.g., 1% or 0.01) or "liberal" (e.g., 10% or 0.1) the investigator wants to be.

So, under the assumption that the null hypothesis is true, the p value indicates that we would expect to see the observed results (or results even more extreme) less than p %

of the time. For example, a p value of 0.02 means that in only 2 occurrences out of a 100 would we expected to see the observed results, if the null hypothesis was true (i.e., that there really is no difference between the groups being tested). Because the p value is lower than our pre-specified significance level or alpha of 0.05, we would reject the null hypothesis and accept the alternative hypothesis.

Note that it is very important to understand what the p values means and as you can imagine the p-value gets misused all the time. Perhaps the most common misinterpretation of the p-value is that it represents the probability of the null hypothesis being true given the data i.e., P (H_0 true|Data). Since, we know that the p-value is calculated under the assumption that the Ho is true, we know that this cannot be the case!

B. Summary - Steps in Classical Hypothesis Testing

1. Define the null hypothesis:

The null hypothesis is that there is no difference between the groups being compared. For example, in a clinical trial the null hypothesis might be that the response rate in the treatment group is equal to that in the control group.

2. Define the alternative hypothesis:

The alternative hypothesis would be that the response rate in the treatment group is different from the control group (two sided test) or is greater (or lesser) than the control group (one sided test).

3. Calculate a *p* value:

This calculation assumes that the null hypothesis is true. One determines the probability of obtaining the results found in the data (or results even more extreme) given the null hypothesis is true. This probability is the *p* value.

4. Accept or reject the null hypothesis:

If the probability of observing the actual data (or more extreme results), under the null hypothesis is smaller than the significance level ($p < \alpha$), then we reject the null. The concept is that if the probability under the null hypothesis of observing the actual results is very small, then there is a conflict between the null hypothesis and the observed data, and we should conclude that the null hypothesis is not true.

5. Accept the alternative hypothesis:

If we reject the null hypothesis, we accept the alternative hypothesis by default. The only way the alternative can be accepted is by rejecting the null (this process of scientific inference is referred to as *refutation*).

C. Summary of Selected Statistical Tests

Depending on the characteristics of the data being analyzed, different statistical tests are used to determine the *p* value. The most common statistical tests are described in the Table below.

Statistical Test	Description
Student's t test	Used to test whether or not the means from two groups using continuous data are equal, assuming that the data are normally distributed and that the data from both groups have equal variance (parametric).
Wilcoxon rank sum test (Mann-Whitney U test)	Used to test whether two sets of observations have the same distribution. These tests are similar in use to the t test, but do not assume the data are normally distributed (non-parametric).
Chi-square test	Used with categorical variables (two or more discrete treatments with two or more discrete outcomes) to test the null hypothesis that there is no effect of treatment on outcome. To be valid the chi-square test requires at least 5 observations in each 'cell" of the cross-classification table.
Fisher's exact test	Used in an analogous manner to the chi-square test, Fisher's exact test may be used even when less than 5 observations in one or more 'cells'.
One-way ANOVA*	Used to test the null hypothesis that three or more sets of continuous data have equal means, assuming the data are normally distributed and that the data from all groups have identical variances (parametric). The one-way ANOVA may be thought of as a t test for three or more groups.
Kruskal-Wallis	This is a non-parametric test analogous to the one-way ANOVA. No assumption is made regarding normality of the data. The Kruskal-Wallis test may be thought of as a Wilcoxon rank sum test for three or more groups.

* Analysis of variance.

Student's t test and Wilcoxon's rank sum test are used to compare continuous variables (e.g., serum glucose, respiratory rate, etc.) between two groups of patients. If there are three or more groups of patients, then one-way analysis of variance (ANOVA) and the Kruskal-Wallis test are used to compare continuous variables between the groups. The chi-square test and Fisher's exact test are used to detect associations when both the treatment and the outcome are categorical variables (placebo versus active drug, lived versus died, admitted versus discharged, etc.). Student's t test and one-way analysis of variance are examples of parametric statistical tests. Parametric statistical tests make assumptions about the underlying distribution of the continuous variables. Both Student's t test and analysis of variance assume that the data are normally distributed, and that the different groups yield data with equal variance.

When the data to be analyzed are not normally distributed, then the *p* value should be obtained using a non-parametric test. Non-parametric tests are referred to as 'distribution-free' in that they do not rely on the data having any particular underlying distribution. The non-parametric alternative to a t test is the Wilcoxon rank sum test or the Mann-Whitney U test. The non-parametric alternative to one-way analysis of variance is the Kruskal-Wallis test.

D. Type I (alpha) and Type II (beta) errors

When we either accept or reject the null hypothesis, there are two types of errors one can make:

A Type I or a FP error

A Type II or a FN error

The relationship between significance testing and the truth can be illustrated in the following 2 x 2 table, which is set up in exactly the same fashion as the classic diagnostic 2 x 2 table (see lecture on clinical testing). However, in this case we are using the significance test (i.e., $P < 0.05$ or $P > 0.05$) as a diagnostic test to infer the truth (i.e., whether the Ho is true or false), just as in the same manner that we use a diagnostic test to infer whether disease is present or not.

		TRUTH	
		H_0 False	H_0 True
SIGNF TEST	REJECT H_o ($P \leq 0.05$)	TP	FP Type I (a)
	ACCEPT H_0 ($P > 0.05$)	FN Type II (B)	TN

(1 – B) (= Power)

Type I (FP) errors

A type I error occurs when we reject the H_0 when it is true – that is we determine a difference exists when it does not (hence it is a type of "false-positive" FP result). A type I error occurs when a statistically significant *p* value is obtained when, in fact, there is no underlying difference between the groups being compared. The rate that FP results are expected to occur is measured by the *significance level* (α) which is also referred to as the *Type I error rate*. As discussed previously this is set, by convention, to 5%.[1] This 5% rejection rule is used primarily because scientists by nature are cautious, so they want a low error rate to avoid false alarms. This is similar to judges providing the instructions to a jury to "only find the person guilty beyond a reasonable doubt". You want to avoid finding someone guilty who is actually innocent.

[1] As will become more apparent after completing the Clinical Testing lecture, setting the alpha level to 0.05 is equivalent to making all significance tests have a specificity of 95%.

Type II (FN) errors

A type II error occurs when we accept the Ho when it is false – that is we determine that a difference does not exists when in fact it does (hence it is a type of "false-negative" FN result). A type II error occurs when a statistically **non-significant** *p* value is obtained when, in fact, there is an underlying difference between the groups being compared. The rate that FN results are expected to occur is measured by the *Type II error rate* (or *beta*). Unlike alpha, beta is not set to any particular level, although, for studies that are actually being planned or designed, sample size estimates are usually based on setting beta to either 20% or even as low as 10%. Thus, such studies are set up with the expectation that a real difference would be missed between 10 and 20% of the time. However, it should be noted that many studies are not based on any formal sample size estimates, and so, particularly for smaller studies, the probability of a type II error maybe a lot higher.

E. Power and Sample Size

The complement of the type II error rate (i.e., 1 - beta) is called Power and is defined as:

Probability of correctly rejecting Ho when Ho is false

Power is therefore the probability of the study finding a difference when a difference truly exists.[2] Power is a function 4 parameters:

i) Alpha (FP) error rate ii)
Beta (FN) error rate iii)
Effect size
iv) The variability in the data

Alpha (FP) error rate

As discussed above, the *significance or alpha level* is usually set at 5%. There is an inverse relationship between alpha and beta – as one increases the other must decrease (and vice versa). So, if alpha is made smaller (for example, reduced from 0.05 to 0.01) the beta error must increase, which would lower the Power of the study making it harder for the study to identify a real difference (a type II error is more likely). This makes sense as a lower alpha is a more stringent test so it is harder to prove that a difference truly exists i.e., it is harder to reject the null hypothesis if the alpha level is changed from 5% to 1%.

Beta (FN) error rate

[2] As will become more apparent after completing the Clinical Testing lecture, Power is equivalent to the Sensitivity of the experiment – the probability of finding a difference if a difference exits is equivalent in diagnostic testing to the probability of finding disease if disease exists (i.e., the test sensitivity)

The smaller the beta error rate the easier it will be for the study to identify a difference (because the Power is increased). The simplest way to increase the Power of a study is to increase the sample size. Large studies have more Power and so for a given treatment effect are more likely to identify a difference as statistically significant. Alternatively, Power can be increased by increasing alpha (say, from 0.05 to 0.10), since beta must decrease. Typically beta is set at 20%, which is four times larger than alpha (this reflects the greater attention placed on FP vs. FN errors – so scientist inherently value FP and FN errors differently – they are so risk adverse that they set the FP rate much lower than the FN rate)

<u>Effect size</u>
The effect size is the magnitude of the treatment difference you are trying to detect. Bigger differences are obviously easier to detect than smaller differences – this is why in statistics <u>size does matter.</u> When designing the study it is important to determine what size of an effect do you want to detect. This is usually determined by defining the *minimal clinically important difference* i.e., what is the smallest difference between 2 alternative treatments that you would want to know from a clinical standpoint? For example, in the trial of sympathomimetic agents for patients with sepsis it might be determined that clinically a 5 mm Hg increase in mean blood pressure is likely to make an important difference to the clinical outcomes of patients. The study would then be powered to be able to detect at least this difference. Sometimes however, a larger minimal treatment effect is defined (e.g., 15 mm Hg), because designing a study to reliably detect the smallest clinically important treatment effect (i.e., 5 mm Hg) would require too large a sample size.

<u>The Variability in the data</u>
The greater the variability in the data the harder it is to detect a difference (the Power is lower). By analogy, it is harder to detect the true "signal" when there is a lot of "noise" to contend with. This effect of variability on study Power is also seen in studies that "count" events. If such outcomes are rare (e.g., death, or relapse in a follow-up study) then the study will have low Power - since it is harder to identify any real differences.

These 4 parameters are all considered when determining the sample size of the study i.e., how big of a study do I need to detect the *minimal clinically important difference*. Most well designed clinical studies are usually set up to have a power of 0.80 (80%) or greater. But it is surprisingly common for clinical studies to be published with negative results, which did not have adequate sample size to reliably detect a clinically important treatment effect in the first place. The problem with low power studies is that it is difficult to interpret negative results. Is there truly no effect?, or did the study simply fail to detect the true effect that does exist because it was too small or had too few outcomes? A low power study also means that any estimate will be measured imprecisely – as indicated by a wider confidence interval.

F. Summary of commonly used terms and definitions in classical hypothesis testing

Term	**Definition**
α (*significance level)*	The maximum *p* value to be considered statistically significant; the risk of committing a type I error when there is no difference between the groups.
Alternative Hypothesis	The hypothesis that is considered as an alternative to the null hypothesis; usually the alternative hypothesis is that there is an effect of the studied treatment on the measured variable of interest; sometimes called the test hypothesis.
β	The risk of committing a type II error.
Null Hypothesis	The hypothesis that there is no effect of the studied treatment on the measured variable of interest.
Power	Probability of correctly rejecting Ho when Ho is false. It is the probability of detecting a treatment effect if one truly exists. Power = 1 – β.
p value	The probability of obtaining results similar (or more extreme) to those actually obtained if the null hypothesis were true.
Type I Error	Obtaining a statistically significant *p* value when, in fact, there is no effect of the studied treatment on the measured variable of interest; a false-positive result.
Type II Error	Not obtaining a statistically significant *p* value when, in fact, there is an effect of the treatment on the measured variable of interest that is as large or larger than the effect the trial was designed to detect; a false-negative result.

II. Estimation, Point Estimates and Confidence Intervals

Estimation is an alternative approach to statistical inference which views experimentation as a measurement exercise. Estimation is concerned with estimating the *specific value* of an unknown population parameter and measuring the precision with which this specific value is measured. Thus, estimation involves the generation of a point estimate and confidence interval (CI), respectively.

A point estimate is the *observed single best estimate of the unknown treatment effect (or population parameter), given the data*. It indicates the *magnitude* of an effect and answers the question: *What did you find or estimate*?

A confidence interval is the *set of all possible values for the parameter that are consistent with the data.* It serves to quantify the *precision* of the estimate and answers the question: *With what sort of precision did you measure the point estimate*?

Suppose we wish to test whether one vasopressor is better than another, based on the mean post-treatment systolic blood pressure (SBP) in hypotensive patients and, in our trial, we observe a mean post-treatment SBP of 70 mm Hg for patients given vasopressor A and 95 mm Hg for patients given vasopressor B. The observed treatment difference or point estimate (mean SBP for patients on vasopressor B minus mean SBP for patients on vasopressor A) is therefore 25 mm Hg. When hypothesis testing, if the *p* value is less than 0.05 we reject the null hypothesis as false and we conclude that our study demonstrates a statistically significant difference in mean SBP between the groups. That the *p* value is less than 0.05 tells us only that the treatment difference that we observed is statistically significantly different from zero. It does not tell us the size of the treatment difference, which determines whether the difference is *clinically* important, nor how precisely our trial was able to estimate the true treatment difference (N.B. The true treatment difference is the difference that would be observed if all similar hypotensive patients could be included in the study).

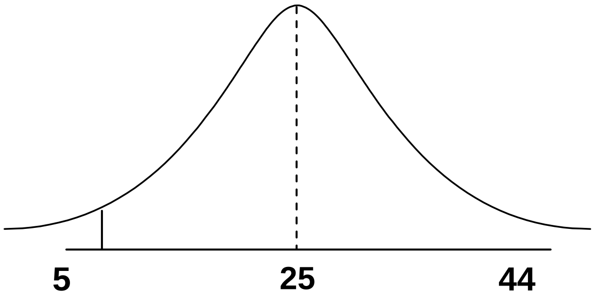
However, if instead of using hypothesis testing and reporting a *p* value, we view the trial's data as a measurement exercise, we would report the point estimate and the corresponding confidence interval surrounding it. This would give readers the same information as the *p* value, plus several other pieces of valuable information including:

- the size of the treatment difference (and therefore its clinical importance),
- the precision of the estimated difference,
- information that aids in the interpretation of a negative result.

The *p* value answers only the question, "Is there a statistically significant difference between the two treatments?" The point estimate and its confidence interval also answer the questions, "What is the size of that treatment difference (and is it clinically important)?" and "How precisely did this trial determine or estimate the true treatment difference?" As clinicians, we should change our practice only if we believe the study has definitively demonstrated a treatment difference, and that the treatment difference is large enough to be clinically important. Even if a trial does not show a statistically significant difference, the confidence interval enables us to distinguish whether there really is no difference between the treatments, or the trial simply did not have enough patients to reliably demonstrate a difference.

Returning to our example, a treatment difference of 0 is equivalent to the null hypothesis that there is no difference in mean SBP between patients on vasopressor A and patients on vasopressor B. In our hypothetical trial, the 95% confidence interval around the point estimate of 25 mm Hg was estimated to be 5 to 44 mm Hg which does not include 0, and so a true treatment difference of zero is not statistically consistent with our data. We therefore conclude that the null hypothesis is not statistically consistent with our observed data and we reject it, accepting the alternative hypothesis. In this example, because the 95% confidence interval does not include a zero treatment difference, this demonstrates that the results are statistically significant at $p < 0.05$.

The confidence interval of 5 to 44 mm Hg with a point estimate of 25 would look like this:



Our point estimate of 25 mm Hg gives an estimate for the size of the treatment difference. However, our results are also statistically consistent with any value within the range of the confidence interval of 5 to 44 mm Hg. In other words, the true treatment difference may be as little as 5 mm Hg, or as much as 44 mm Hg, however 25 mm Hg remains the most likely value given our data. If vasopressor B has many more severe side effects than vasopressor A, a reader may conclude that even an elevation of SBP as much as 44 mm Hg does not warrant use of vasopressor B, although the treatment difference is statistically significant. Another reader may feel that even an elevation in mean SBP of 5 mm Hg would be beneficial, despite the side effects. With *p* values, authors report a result as statistically significant or not, leaving us with little basis for drawing conclusions relevant to our clinical practice. With confidence intervals we may decide what treatment difference we consider to be clinically important, and reach conclusions appropriate for our practice.

We may also use confidence intervals to obtain important information from trials that did not achieve statistical significance (so called “negative” trials). Suppose we found the 95% confidence interval for the difference in mean SBP to be -5 mm Hg to 55 mm Hg, with the same point estimate of 25 mm Hg. Now our results are consistent with vasopressor B raising mean SBP as much as 55 mm Hg *more* than vasopressor A, or as much as 5 mm Hg *less*. Because the confidence interval includes 0 (a zero treatment difference equivalent to the null hypothesis), the results are not statistically significant and the p value is > 0.05. Since $p > 0.05$, we may be tempted to conclude that there is no advantage to using vasopressor A or B in our clinical practice. However, our data are still consistent with vasopressor B raising SBP by 25 mm Hg (i.e., the point estimate) and the data are also consistent with as much as a 55 mm Hg increase when vasopressor B is used. Although $p > 0.05$, there remains the possibility that a clinically important difference exists in the two vasopressors' effects on mean SBP. Negative trials whose results are still consistent with a clinically important difference usually occur when the sample size was too small, resulting in low power to detect an important treatment difference.

It is important to know how precisely the point estimate represents the true difference between the groups. The width of the confidence interval gives us information on the precision of the point estimate. The larger the sample size the more precise the point estimate will be, and the confidence interval will be narrower. As mentioned above, negative trials that use too small a sample size may often not show a statistically significant result, yet still not be able to exclude a clinically important treatment difference. In this case, the confidence interval is wide and imprecise, and includes both zero (or no treatment difference), as well as clinically important treatment differences. Conversely, positive trials that use a very large sample size may show a statistically significant treatment difference that is not clinically important, for example an increase in mean SBP from 70 mm Hg to 72 mm Hg.

If a confidence interval includes both zero as well as clinically important treatment differences, we can not make any definitive conclusions regarding clinical practice. It is important to remember that the data are statistically consistent with the true value being anywhere within the entire range of the confidence interval.

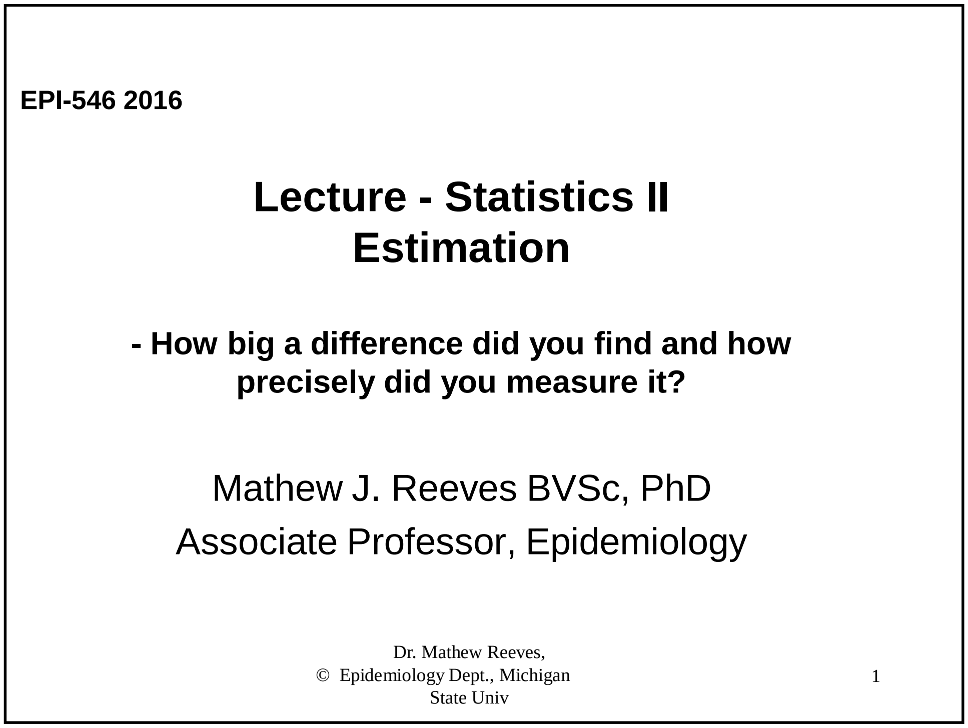
III. Multiple Comparisons Whenever two groups of patients are compared statistically, even if they are fundamentally identical, there is a chance that a statistically significant *p* value will be obtained. If the significant level (α) is set to 0.05, then there is a 5% chance that a statistically significant *p* value will be obtained even if there is no true difference between the two patient populations (remember the p value is defined on the basis of there being no difference between the null and alternative hypotheses). The risk of a false-positive *p* value occurs each time a statistical test is performed. When *multiple* comparisons are performed, for example, the comparison of many different characteristics between two groups of patients (i.e., sub-group analyses), the risk of at least one false-positive *p* value is increased, because the risk associated with each test is incurred multiple times. The risk of obtaining at least one false-positive *p* value, when comparing two groups of fundamentally identical patients, is shown in Table 4 as a function of the number of independent comparisons (i.e., statistical tests) made. For up to 5 to 10 comparisons, the overall risk of at least one type I error is roughly equal to the maximum significant *p* value used for each individual test, multiplied by the total number of tests performed.

Number of Comparisons	**Probability of at Least One Type I Error**
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
10	0.40
20	0.64
30	0.79

The Bonferroni correction is a method for reducing the overall type I error risk for the whole study by reducing the maximum *p* value used for each of the individual statistical tests. The overall risk of a type I error that is desired (usually 0.05) is divided by the number of statistical tests to be performed, and this value is used as the maximum significant *p* value for each individual test. For example, if five comparisons are to be made, then a maximum significant *P* value of 0.01 should be used for each of the five statistical tests.

The Bonferroni correction controls the overall (study-wise) risk of a type I error, at the expense of an increased risk of a type II error. Since each statistical test is conducted using a more stringent criteria for a statistically significant *p* value, there is an increased risk that each test might miss a clinically-important difference, and yield a *p* value that is non-significant using the new criteria of $p < 0.01$. The Bonferroni correction is regarded as a very conservative approach to the problem of multiple comparisons.

IV. Multivariable analysis and interaction - (this topic is covered in EPI-547 the second year course).

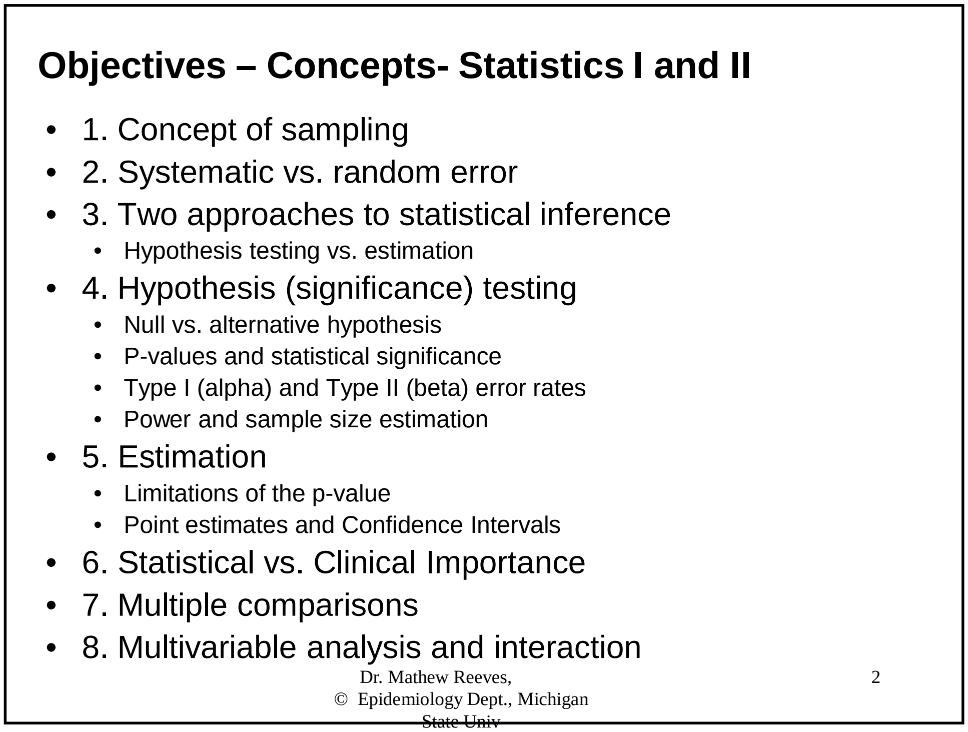


EPI-546 2016

Lecture - Statistics II
Estimation

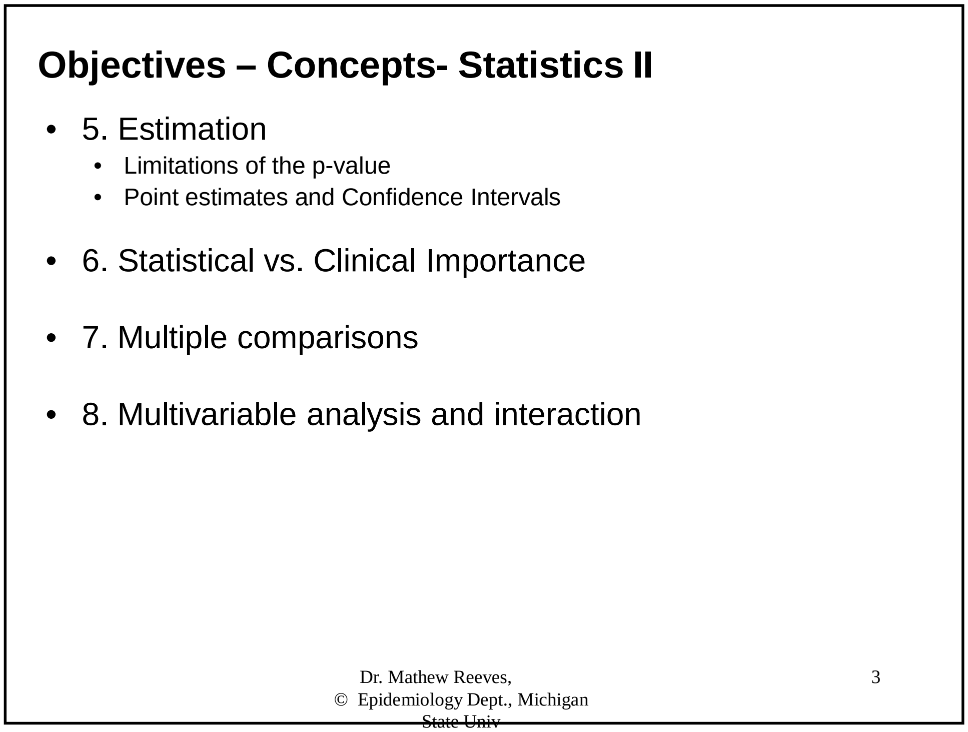
- How big a difference did you find and how precisely did you measure it?

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology



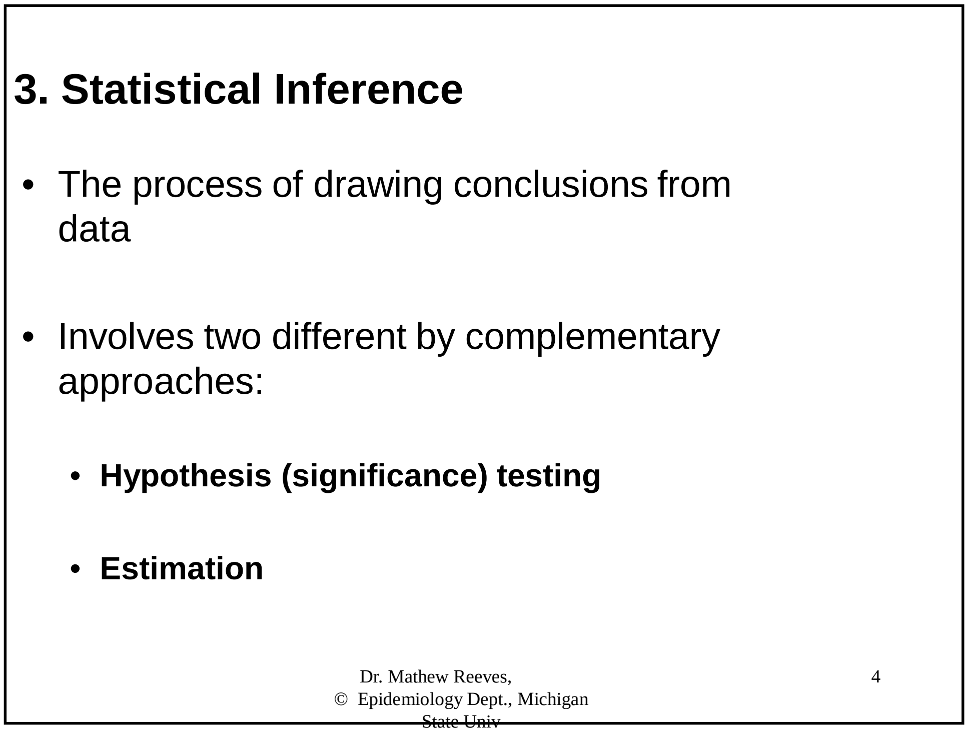
Objectives – Concepts- Statistics I and II

- 1. Concept of sampling
- 2. Systematic vs. random error
- 3. Two approaches to statistical inference
  - Hypothesis testing vs. estimation
- 4. Hypothesis (significance) testing
  - Null vs. alternative hypothesis
  - P-values and statistical significance
  - Type I (alpha) and Type II (beta) error rates
  - Power and sample size estimation
- 5. Estimation
  - Limitations of the p-value
  - Point estimates and Confidence Intervals
- 6. Statistical vs. Clinical Importance
- 7. Multiple comparisons
- 8. Multivariable analysis and interaction



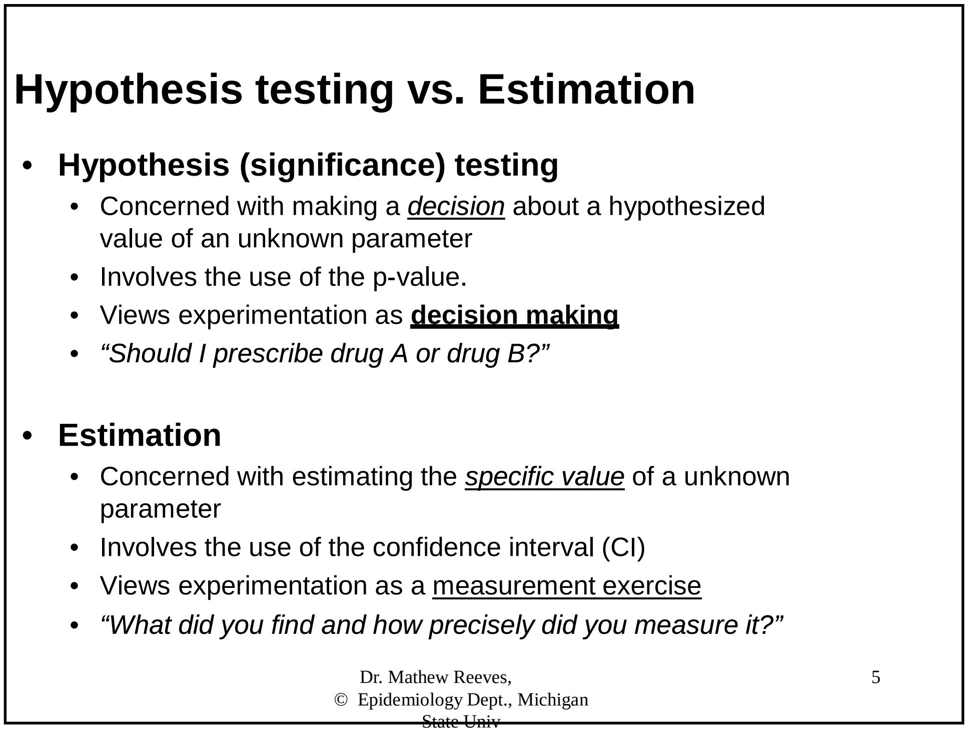
Objectives – Concepts- Statistics II

- 5. Estimation
  - Limitations of the p-value
  - Point estimates and Confidence Intervals
- 6. Statistical vs. Clinical Importance
- 7. Multiple comparisons
- 8. Multivariable analysis and interaction



3. Statistical Inference

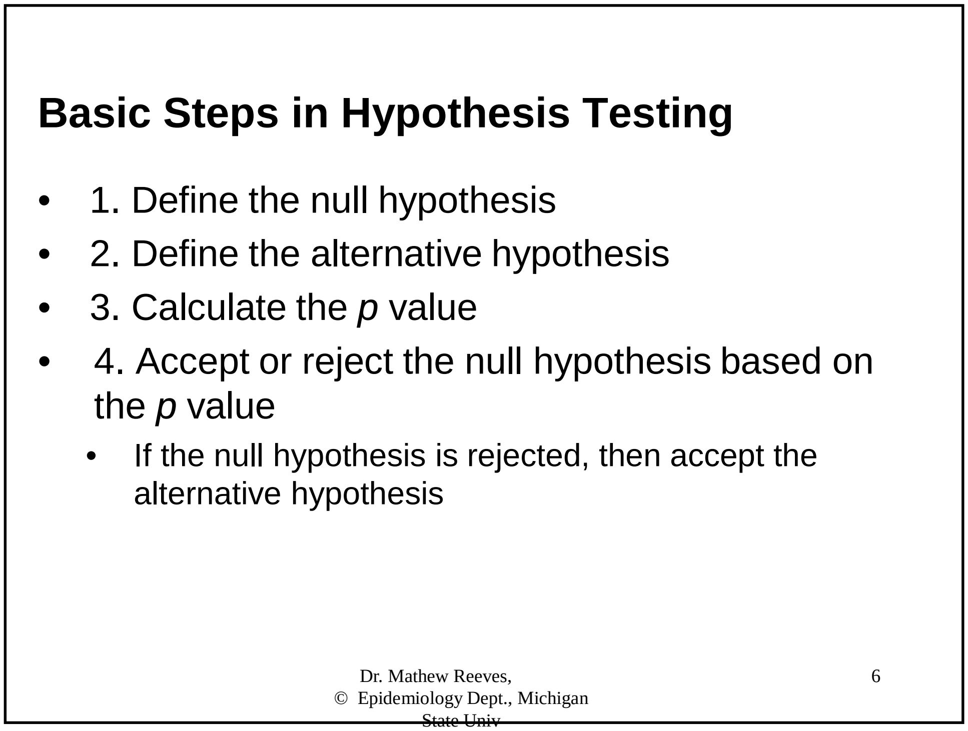
- The process of drawing conclusions from data
- Involves two different by complementary approaches:
  - **Hypothesis (significance) testing**
  - **Estimation**



Hypothesis testing vs. Estimation

- **Hypothesis (significance) testing**
  - Concerned with making a *decision* about a hypothesized value of an unknown parameter
  - Involves the use of the p-value.
  - Views experimentation as **decision making**
  - *"Should I prescribe drug A or drug B?"*

- **Estimation**
  - Concerned with estimating the *specific value* of a unknown parameter
  - Involves the use of the confidence interval (CI)
  - Views experimentation as a measurement exercise
  - *"What did you find and how precisely did you measure it?"*



Basic Steps in Hypothesis Testing

- 1. Define the null hypothesis
- 2. Define the alternative hypothesis
- 3. Calculate the *p* value
- 4. Accept or reject the null hypothesis based on the *p* value
  - If the null hypothesis is rejected, then accept the alternative hypothesis

Clinical Study (statistical testing)	Jury Trial (criminal law)
Assume the null hypothesis	Presume innocent
Goal: detect a true difference (reject the null hypothesis)	Goal: convict the guilty
"Level of significance" p < .05	"Beyond reasonable doubt"
Requires: adequate sample size	Requires: convincing testimony

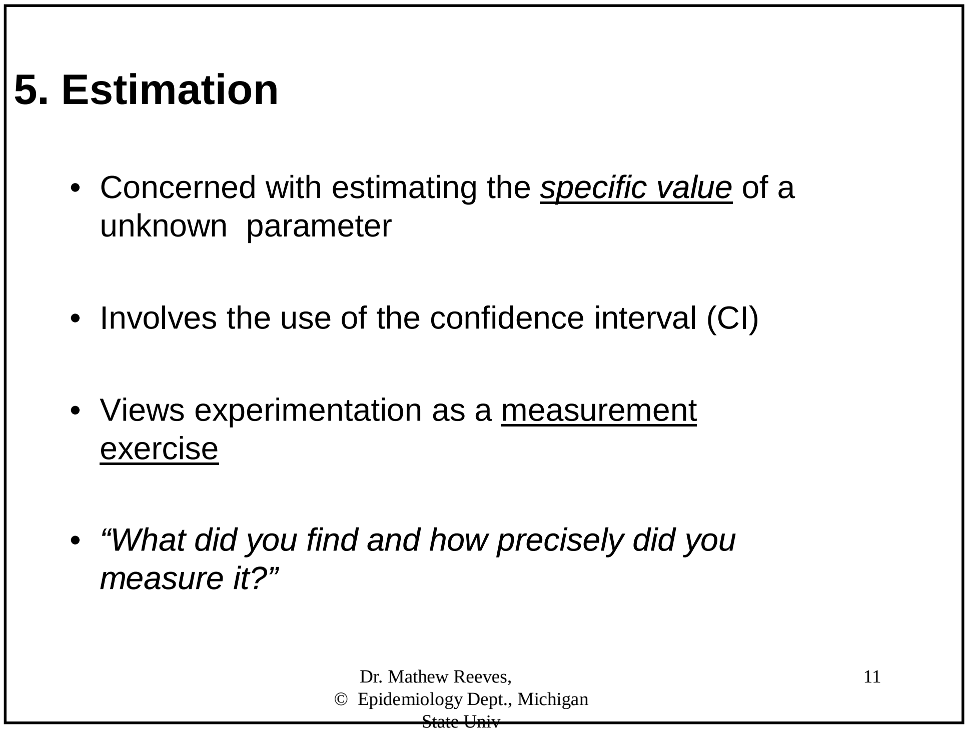
Similar to clinical study in a Trial by Jury…..

- There are only 1 of 4 possible outcomes:
  - 2 are correct: TP, TN
  - 2 are errors: FP, FN

Trial by Jury…..

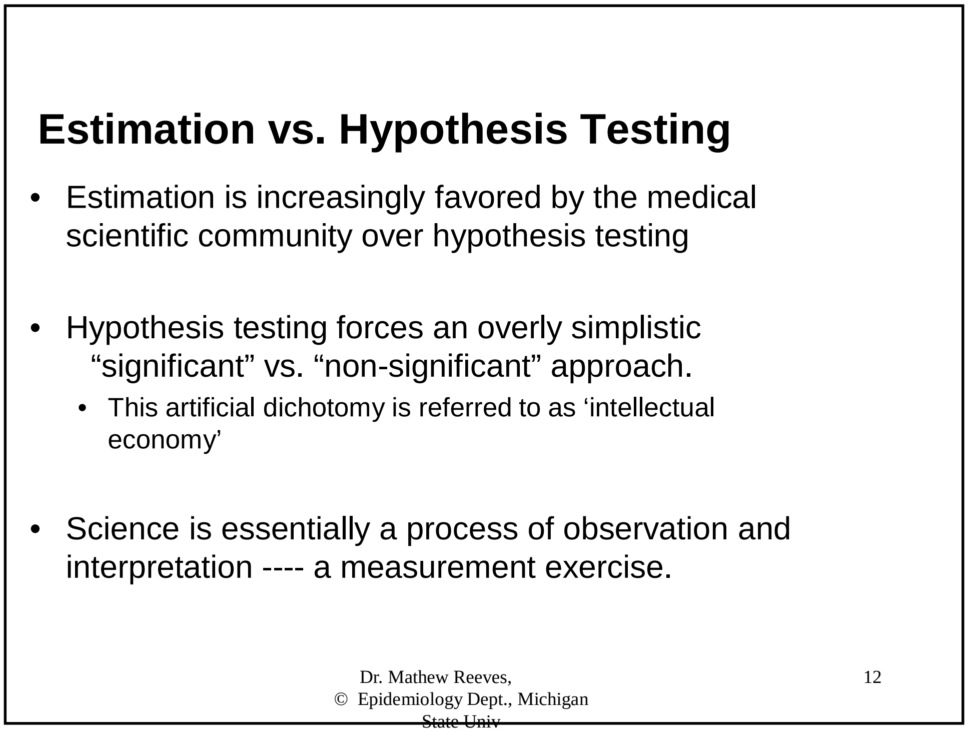
		TRUTH	
		Guilty	Innocent
Jury Decision	Guilty	TP	FP
	Not guilty	FN	TN

Clinical Study (statistical testing)	Jury Trial (criminal law)
Correct inference: reject the null hypothesis	Correct verdict: convict a guilty person
Correct inference: accept the null hypothesis	Correct verdict: acquit the innocent
Incorrect inference (FP) Type I error	Incorrect verdict: hang innocent person
Incorrect inference (FN) Type II error	Incorrect verdict: guilty goes free



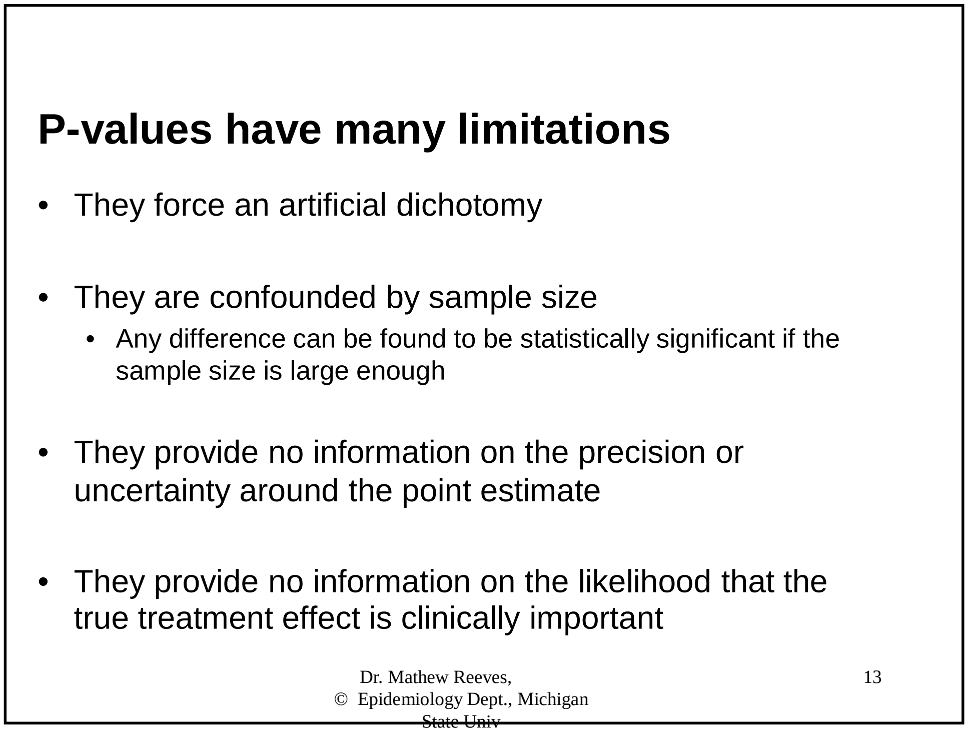
5. Estimation

- Concerned with estimating the *specific value* of a unknown parameter
- Involves the use of the confidence interval (CI)
- Views experimentation as a measurement exercise
- *“What did you find and how precisely did you measure it?”*



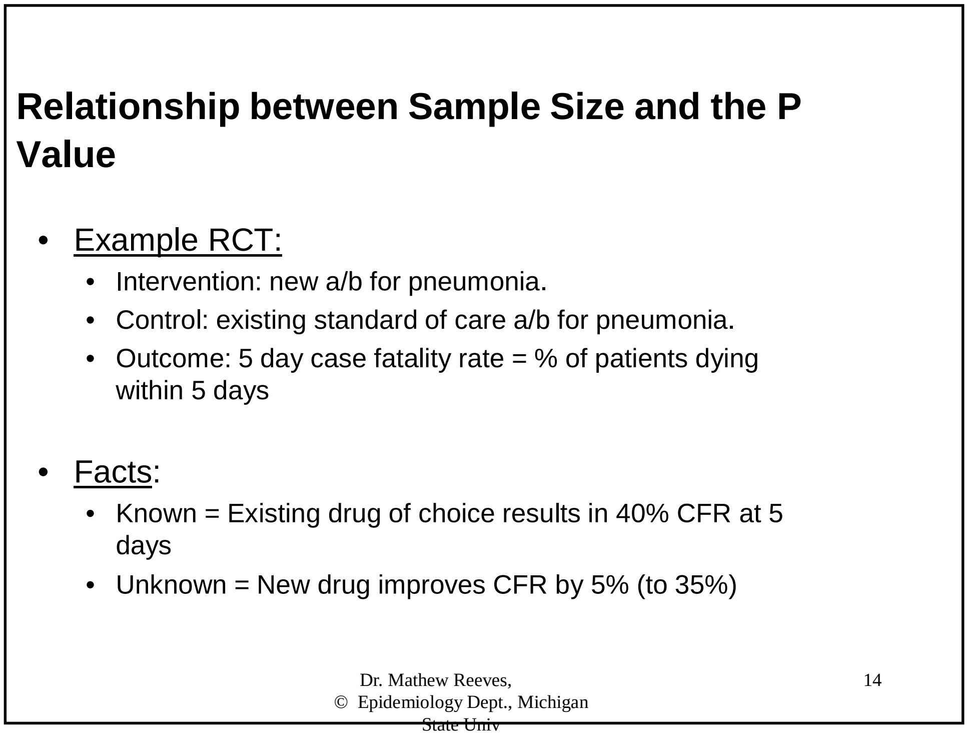
Estimation vs. Hypothesis Testing

- Estimation is increasingly favored by the medical scientific community over hypothesis testing
- Hypothesis testing forces an overly simplistic “significant” vs. “non-significant” approach.
  - This artificial dichotomy is referred to as ‘intellectual economy’
- Science is essentially a process of observation and interpretation ---- a measurement exercise.



P-values have many limitations

- They force an artificial dichotomy
- They are confounded by sample size
  - Any difference can be found to be statistically significant if the sample size is large enough
- They provide no information on the precision or uncertainty around the point estimate
- They provide no information on the likelihood that the true treatment effect is clinically important



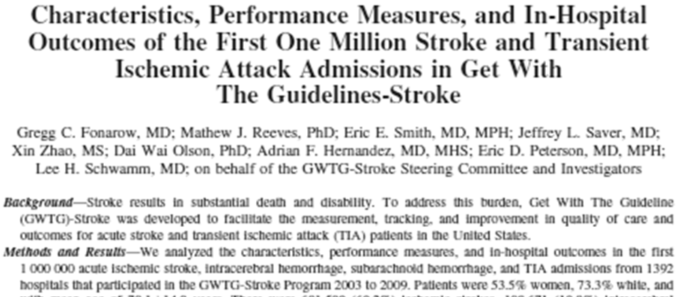
Relationship between Sample Size and the P Value

- Example RCT:
  - Intervention: new a/b for pneumonia.
  - Control: existing standard of care a/b for pneumonia.
  - Outcome: 5 day case fatality rate = % of patients dying within 5 days
- Facts:
  - Known = Existing drug of choice results in 40% CFR at 5 days
  - Unknown = New drug improves CFR by 5% (to 35%)

P-values Generated by RCTs of Different Sample Size For a Constant 5% ARR

Sample Size (N = 2x)	P value (Chi-square)
100 per group	0.465
500	0.103
600	0.074
700	0.053
800	0.039
1000	0.021
2000	<0.001

The problem of too many observations



Characteristics, Performance Measures, and In-Hospital Outcomes of the First One Million Stroke and Transient Ischemic Attack Admissions in Get With The Guidelines-Stroke

Gregg C. Fonarow, MD; Mathew J. Reeves, PhD; Eric E. Smith, MD, MPH; Jeffrey L. Saver, MD; Xin Zhao, MS; Dai Wai Olson, PhD; Adrian F. Hernandez, MD, MHS; Eric D. Peterson, MD, MPH; Lee H. Schwamm, MD; on behalf of the GWTG-Stroke Steering Committee and Investigators

Background—Stroke results in substantial death and disability. To address this burden, Get With The Guideline (GWTG)-Stroke was developed to facilitate the measurement, tracking, and improvement in quality of care and outcomes for acute stroke and transient ischemic attack (TIA) patients in the United States.

Methods and Results—We analyzed the characteristics, performance measures, and in-hospital outcomes in the first 1 000 000 acute ischemic stroke, intracerebral hemorrhage, subarachnoid hemorrhage, and TIA admissions from 1392 hospitals that participated in the GWTG-Stroke Program 2003 to 2009. Patients were 53.5% women, 73.3% white, and

Ref: Circ-CQO 2010
Includes data on One million stroke patients

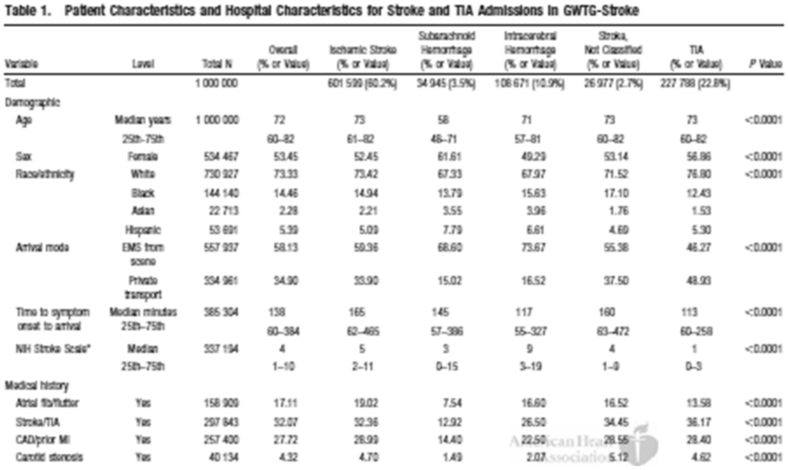


Table 1. Patient Characteristics and Hospital Characteristics for Stroke and TIA Admissions in GWTG-Stroke

Variable	Level	Total N	Overall (% or Value)	Ischemic Stroke (% or Value)	Subarachnoid Hemorrhage (% or Value)	Intracerebral Hemorrhage (% or Value)	Stroke, Not Classified (% or Value)	TIA (% or Value)	*P* Value
Total		1 000 000		601 599 (60.2%)	34 945 (3.5%)	108 671 (10.9%)	26 977 (2.7%)	227 788 (22.8%)	
Demographic									
Age	Median years	1 000 000	72	73	58	71	73	73	<0.0001
	25th–75th		60–82	61–82	46–71	57–81	60–82	60–82	
Sex	Female	534 467	53.45	52.45	61.61	49.29	53.14	56.86	<0.0001
Race/ethnicity	White	730 927	73.33	73.42	67.33	67.97	71.52	76.80	<0.0001
	Black	144 140	14.46	14.94	13.79	15.63	17.10	12.43	
	Asian	22 713	2.28	2.21	3.55	3.96	1.76	1.53	
	Hispanic	53 691	5.39	5.09	7.79	6.61	4.69	5.30	
Arrival mode	EMS from scene	557 937	58.13	59.36	68.60	73.67	55.38	46.27	<0.0001
	Private transport	334 961	34.90	33.90	15.02	16.52	37.50	48.93	
Time to symptom onset to arrival	Median minutes	385 304	138	165	145	117	160	113	<0.0001
	25th–75th		60–384	62–465	57–386	55–327	63–472	60–258	
NIH Stroke Scale*	Median	337 194	4	5	3	9	4	1	<0.0001
	25th–75th		1–10	2–11	0–15	3–19	1–9	0–3	
Medical history									
Atrial fib/flutter	Yes	158 909	17.11	19.02	7.54	16.60	16.52	13.58	<0.0001
Stroke/TIA	Yes	297 843	32.07	32.36	12.92	26.50	34.45	36.17	<0.0001
CAD/prior MI	Yes	257 400	27.72	28.99	14.40	22.50	28.56	28.40	<0.0001
Carotid stenosis	Yes	40 134	4.32	4.70	1.49	2.07	5.12	4.62	<0.0001

The problem is that every difference tested no matter how small is statistically significant P<0.0001!!!

~~P-values~~

Estimation - Example

- Want to estimate the average weight of MSU undergraduates - we want to know the true population mean body weight (μ).

- Take Sample of 20 students:
  - calculate sample mean = 60 kg (= *point estimate*),
  - calculate standard deviation (σ) = 10 kg.
    - (Standard deviation = a measure of variability or dispersion see lecture 2)

- Question:
  - How good an approximation is this mean to the true unknown population mean (μ)?

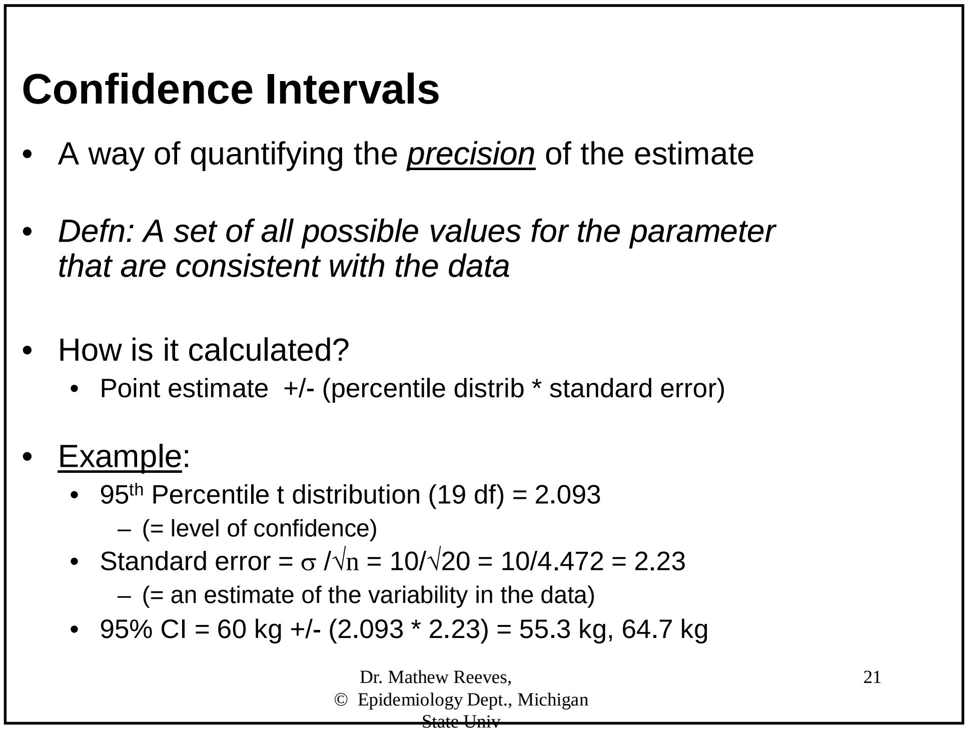
Point estimate

- *Defn: The observed single best estimate of the unknown population parameter (or effect size), given the data*
  - e.g., mean body wt. = 60 kg

- Indicates the *<u>magnitude</u>* of an effect

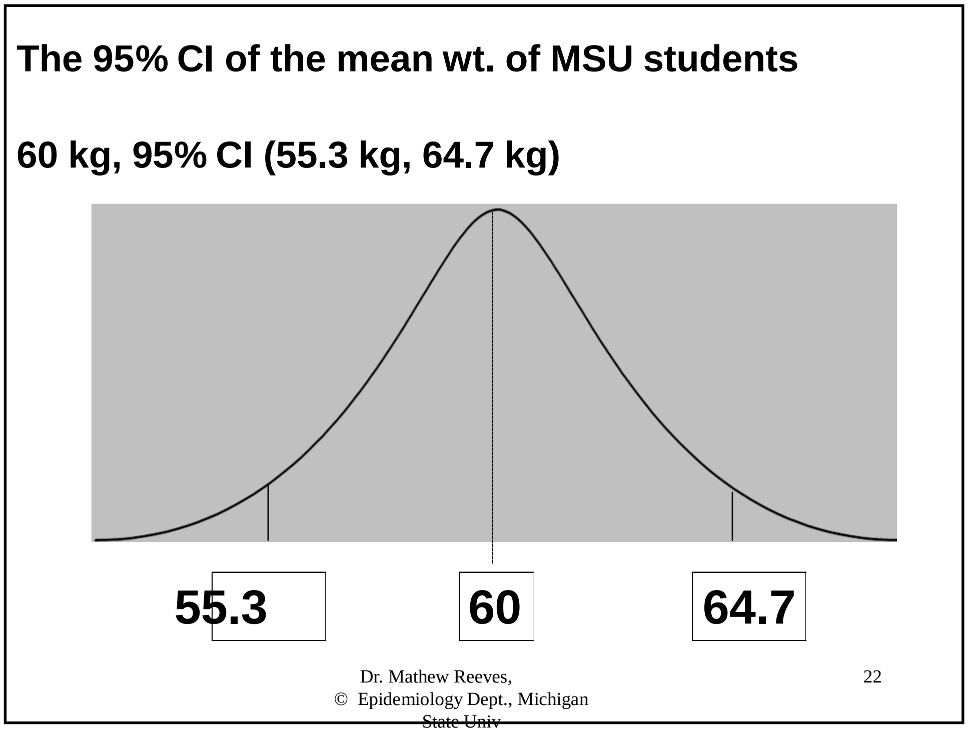
- Answers the question: *What did you find or estimate*?

- But then we also want to know how <u>precisely</u> you were able to measure this?…..



Confidence Intervals

- A way of quantifying the <u>*precision*</u> of the estimate
- *Defn: A set of all possible values for the parameter that are consistent with the data*
- How is it calculated?
  - Point estimate +/- (percentile distrib * standard error)
- <u>Example</u>:
  - 95th Percentile t distribution (19 df) = 2.093
    - (= level of confidence)
  - Standard error = $\sigma / \sqrt{n}$ = 10/√20 = 10/4.472 = 2.23
    - (= an estimate of the variability in the data)
  - 95% CI = 60 kg +/- (2.093 * 2.23) = 55.3 kg, 64.7 kg



The 95% CI of the mean wt. of MSU students

60 kg, 95% CI (55.3 kg, 64.7 kg)

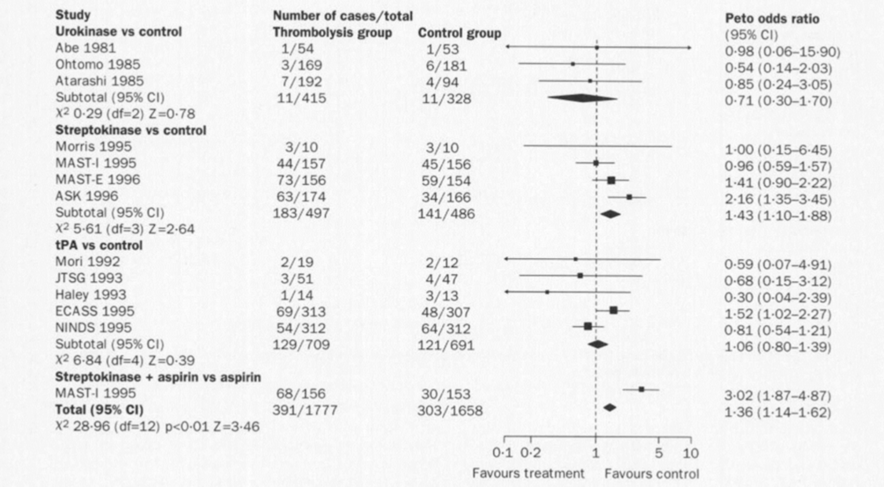
95% CI…… Interpretation

- There is a 95% probability that the true (unknown) mean body weight of all MSU students is included in the interval, 55.3 to 64.7 kg (and our best guess is that it is 60 kg)
- Other points
  - A CI is not a uniform distribution….. values closer to the point estimate (60 Kg) are much more likely than values at the extreme (55.3 and 64.7) (see Figure).
  - CIs can be calculated for any level of confidence… 90%, 99% etc
- Qu: Which is wider a 99% CI or a 90% CI?
  - 99% CI = 60 kg +/- (**2.861** * 2.23) = 53.6 kg, 66.4 kg

Link between confidence intervals and significance testing

- Point estimate (95% CI) = 60 kg (55.3, 64.7)
  - So, all values outside of the 95% CI (i.e., < 55.3 or > 64.7) would be statistically significant from the point estimate of 60 kg at p <0.05.
- Similarly, all values outside of the 99% CI (i.e., < 53.6 or > 66.4) would be statistically significant from the point estimate of 60 kg at p <0.01.

Systematic review of evidence on thrombolytic therapy for acute ischemic stroke
Wardlow JM. Lancet 1997, 350 (607-614)

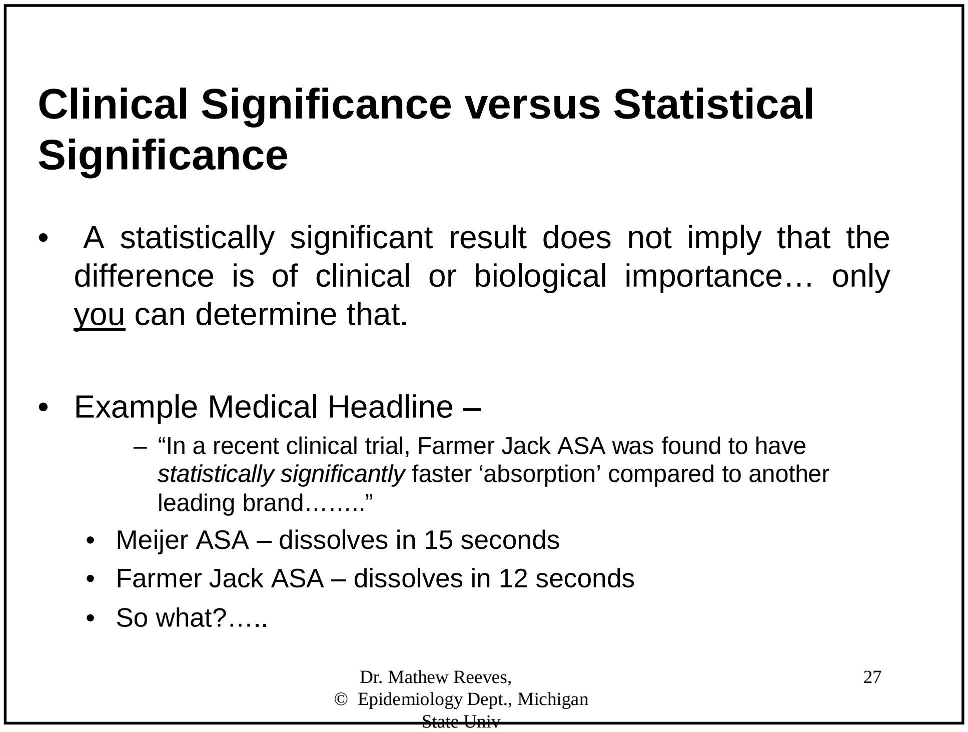


Study	Number of cases/total Thrombolysis group	Control group	Peto odds ratio (95% CI)
Urokinase vs control			
Abe 1981	1/54	1/53	0·98 (0·06–15·90)
Ohtomo 1985	3/169	6/181	0·54 (0·14–2·03)
Atarashi 1985	7/192	4/94	0·85 (0·24–3·05)
Subtotal (95% CI)	11/415	11/328	0·71 (0·30–1·70)
χ^2 0·29 (df=2) Z=0·78			
Streptokinase vs control			
Morris 1995	3/10	3/10	1·00 (0·15–6·45)
MAST-I 1995	44/157	45/156	0·96 (0·59–1·57)
MAST-E 1996	73/156	59/154	1·41 (0·90–2·22)
ASK 1996	63/174	34/166	2·16 (1·35–3·45)
Subtotal (95% CI)	183/497	141/486	1·43 (1·10–1·88)
χ^2 5·61 (df=3) Z=2·64			
tPA vs control			
Mori 1992	2/19	2/12	0·59 (0·07–4·91)
JTSG 1993	3/51	4/47	0·68 (0·15–3·12)
Haley 1993	1/14	3/13	0·30 (0·04–2·39)
ECASS 1995	69/313	48/307	1·52 (1·02–2·27)
NINDS 1995	54/312	64/312	0·81 (0·54–1·21)
Subtotal (95% CI)	129/709	121/691	1·06 (0·80–1·39)
χ^2 6·84 (df=4) Z=0·39			
Streptokinase + aspirin vs aspirin			
MAST-I 1995	68/156	30/153	3·02 (1·87–4·87)
Total (95% CI)	391/1777	303/1658	1·36 (1·14–1·62)
χ^2 28·96 (df=12) p<0·01 Z=3·46			

Figure 3: **Effect of thrombolysis on total case fatality (early and late) at end of trial follow-up**

Estimation - Summary

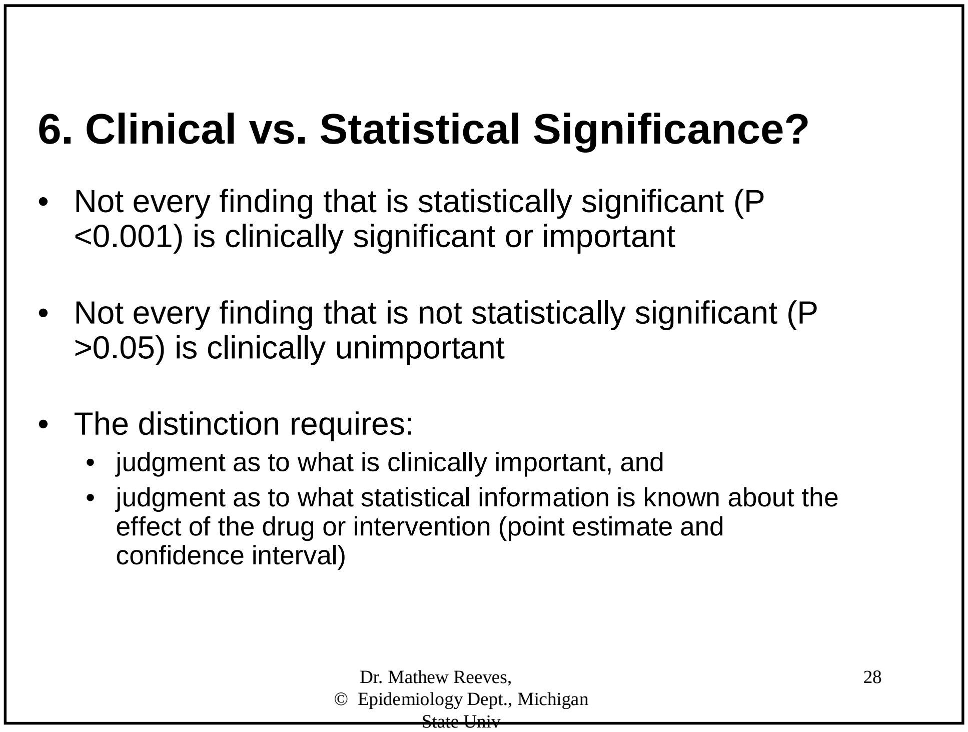
- Views "experimentation" as a <u>measurement exercise</u>
- A *Point Estimate* in conjunction with a *Confidence Interval* is a very powerful combination:
  - Together they indicate the <u>magnitude</u> and <u>precision</u> of the findings
  - Together they answer the question:
    - “What did you find and how precise was your measurement?” (…so how confident are you in your conclusions?)
  - Isn't this all we really need to know?…….



Clinical Significance versus Statistical Significance

- A statistically significant result does not imply that the difference is of clinical or biological importance… only <u>you</u> can determine that.

- Example Medical Headline –
  - – "In a recent clinical trial, Farmer Jack ASA was found to have *statistically significantly* faster 'absorption' compared to another leading brand……."
  - Meijer ASA – dissolves in 15 seconds
  - Farmer Jack ASA – dissolves in 12 seconds
  - So what?…..

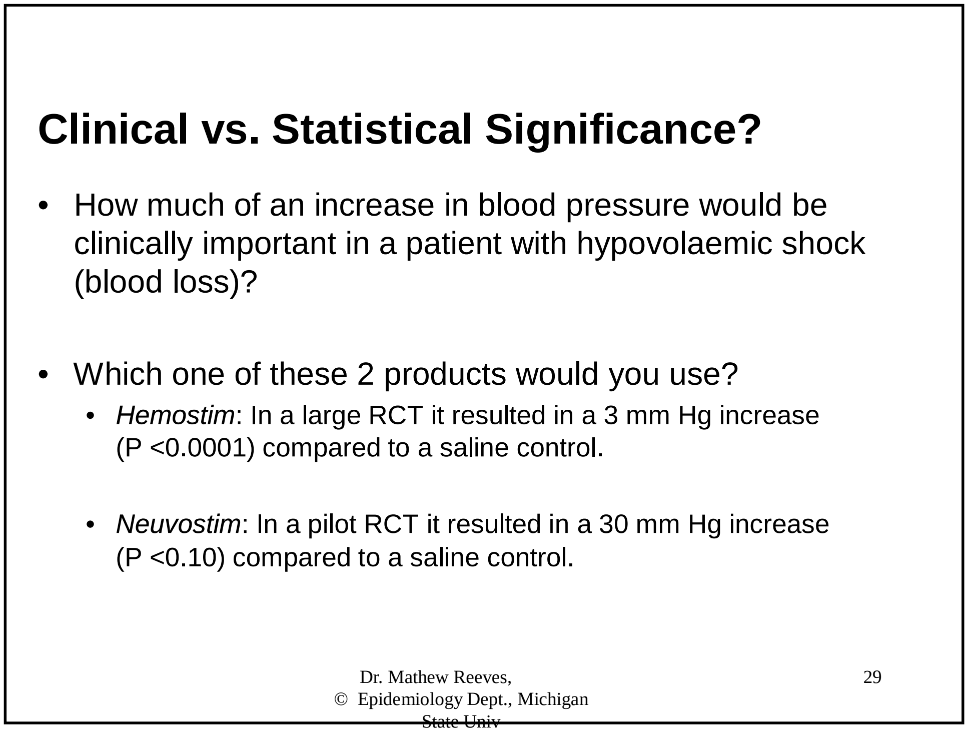


6. Clinical vs. Statistical Significance?

- Not every finding that is statistically significant (P <0.001) is clinically significant or important

- Not every finding that is not statistically significant (P >0.05) is clinically unimportant

- The distinction requires:
  - judgment as to what is clinically important, and
  - judgment as to what statistical information is known about the effect of the drug or intervention (point estimate and confidence interval)

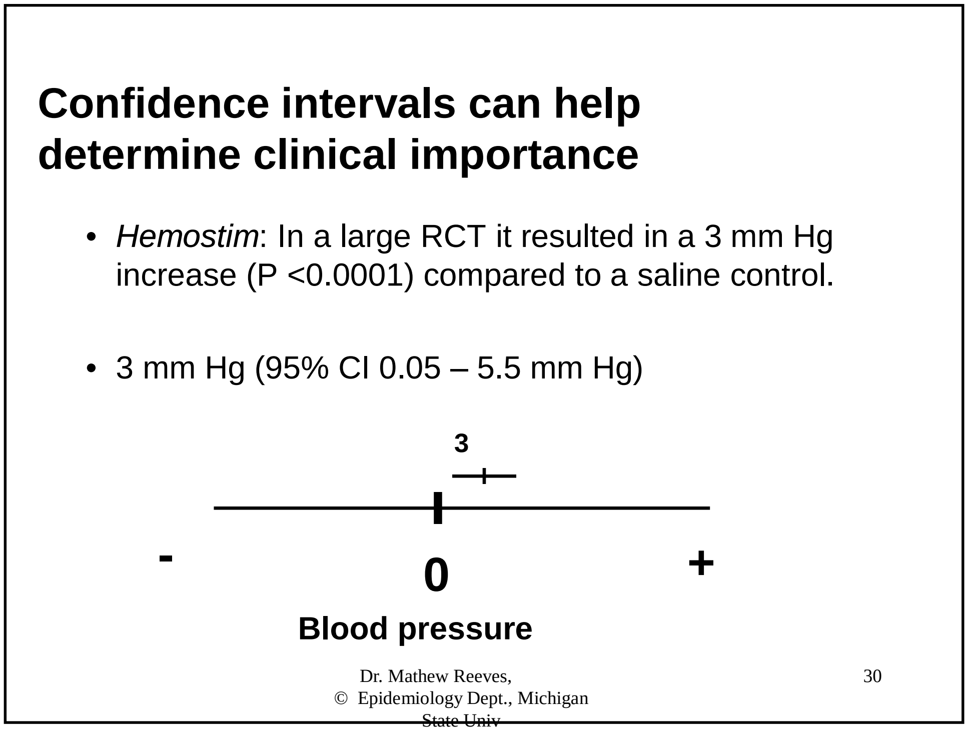


Clinical vs. Statistical Significance?

- How much of an increase in blood pressure would be clinically important in a patient with hypovolaemic shock (blood loss)?

- Which one of these 2 products would you use?
  - *Hemostim*: In a large RCT it resulted in a 3 mm Hg increase (P <0.0001) compared to a saline control.

  - *Neuvostim*: In a pilot RCT it resulted in a 30 mm Hg increase (P <0.10) compared to a saline control.



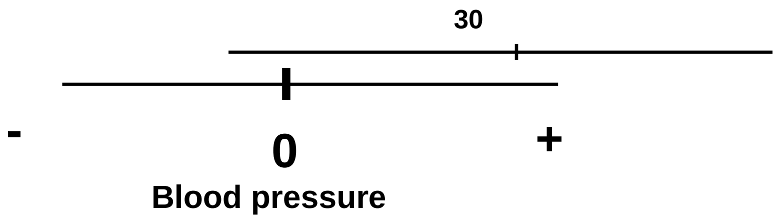
Confidence intervals can help determine clinical importance

- *Hemostim*: In a large RCT it resulted in a 3 mm Hg increase (P <0.0001) compared to a saline control.

- 3 mm Hg (95% CI 0.05 – 5.5 mm Hg)

Confidence intervals can help determine clinical importance

- *Neuvostim*: In a pilot RCT resulted in a 30 mm Hg increase (P <0.10) compared to a saline control.
- 30 mm Hg (95% CI -5.0 – 50 mm Hg)



7. The Problem of Multiple Comparisons

- As with diagnostic tests, the more tests you run the more you are likely to find a significant (abnormal) result due to chance alone
- When multiple comparisons are performed, the risk of one or more false-positive *p* values is increases
- Probability of observing 1 or more “positive” tests if alpha = 0.05
  - Probability = 0.23 when n= 5, and 0.99 when n = 100
    - Where n = number of tests
  - Calculated by [1 – (1- alpha)^n]
  - [1 – (1-0.05)^5] = 0.23

The Problem of Multiple Comparisons

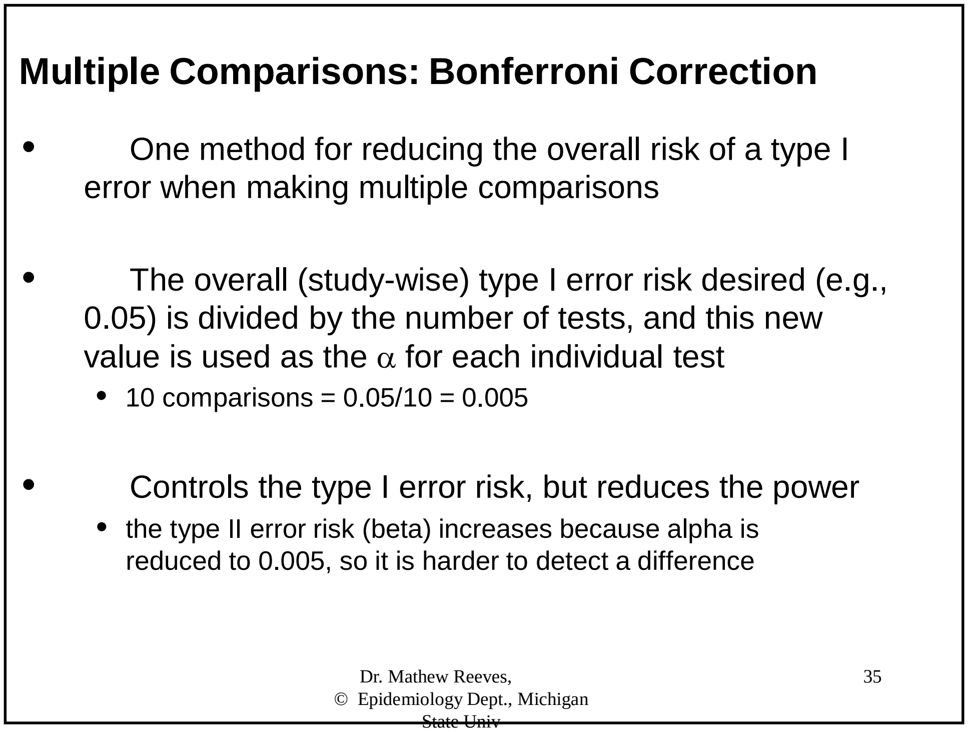
- Certain type of studies are prone to the multiple comparison problem…….
  - Studies that collect a lot of data (large number of variables, and outcome variables)
  - Fishing trips! – not sure what you are looking for
  - Whose *a priori* hypotheses were negative
    - – got to find something “significant” to get it published!
    - – analyze data from many sub-groups (*post-hoc* analyses)

- Multiple comparisons include:
  - – Pair-wise comparisons of more than two groups
  - – The comparison of multiple characteristics between two groups (e.g., sub-group analyses in RCT’s)
  - – The comparison of two groups at multiple time points

Multiple Comparisons: Risk of ≥ 1 False Positive

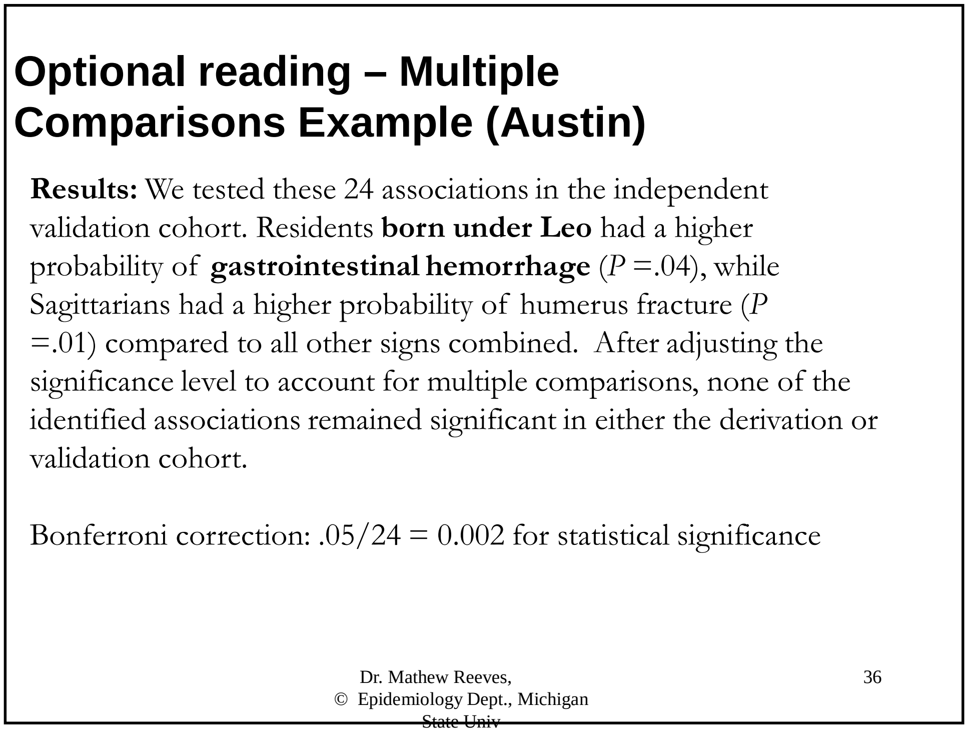
Number of Comparisons	Probability of at Least One Type I Error
1	0.05
2	0.10
3	0.14
4	0.19
5	0.23
10	0.40
20	0.64
30	0.79

Assumes a= 0.05, independent comparisons



Multiple Comparisons: Bonferroni Correction

- One method for reducing the overall risk of a type I error when making multiple comparisons
- The overall (study-wise) type I error risk desired (e.g., 0.05) is divided by the number of tests, and this new value is used as the α for each individual test
  - 10 comparisons = 0.05/10 = 0.005
- Controls the type I error risk, but reduces the power
  - the type II error risk (beta) increases because alpha is reduced to 0.005, so it is harder to detect a difference



Optional reading – Multiple Comparisons Example (Austin)

Results: We tested these 24 associations in the independent validation cohort. Residents **born under Leo** had a higher probability of **gastrointestinal hemorrhage** (*P* =.04), while Sagittarians had a higher probability of humerus fracture (*P* =.01) compared to all other signs combined. After adjusting the significance level to account for multiple comparisons, none of the identified associations remained significant in either the derivation or validation cohort.

Bonferroni correction: .05/24 = 0.002 for statistical significance

8. Multivariable analysis and interaction

- See EPI-547 course notes (second year course).

Finally, statistical Issues to consider when planning a study

- Define the most important question to be answered – the “primary objective”
- Define the size of the difference you wish to detect (minimal clinically important difference)
- Get as much information as possible about what you expect to see in the control group

Statistical Issues to consider when planning a study

- Define values for α and power, and the maximum sample size that is realistic
- Define clinically important subgroups of the population (*a priori* sub-group analyses)
- Determine whether there are important multiple comparisons

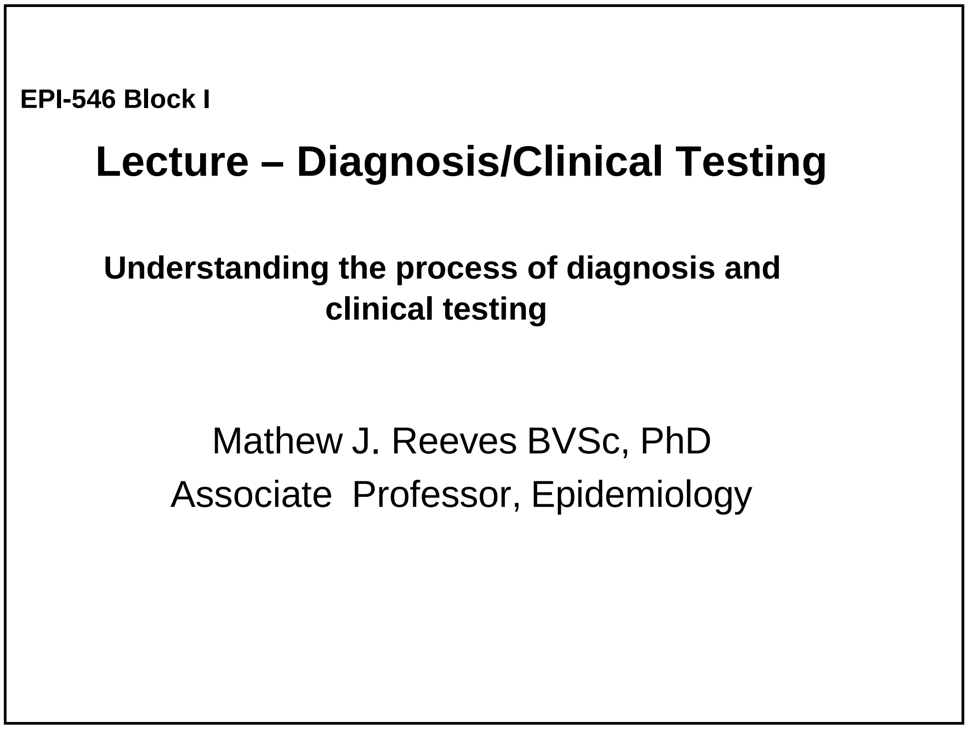
One last example (test of understanding): Significance Testing vs. Interval Estimation

	OUTCOME			
	+	-		
TRT A	7	13	20	P(+ outcome)= 35%
TRT B	14	6	20	P (+ outcome)= 70%

Significance test (TRT A vs. TRT B): P= 0.06 or NS!

Interval estimation: ARR = 35% (95%CI = -1%, +71%)

N.B. This CI illustrates that the difference is NS because it includes 0

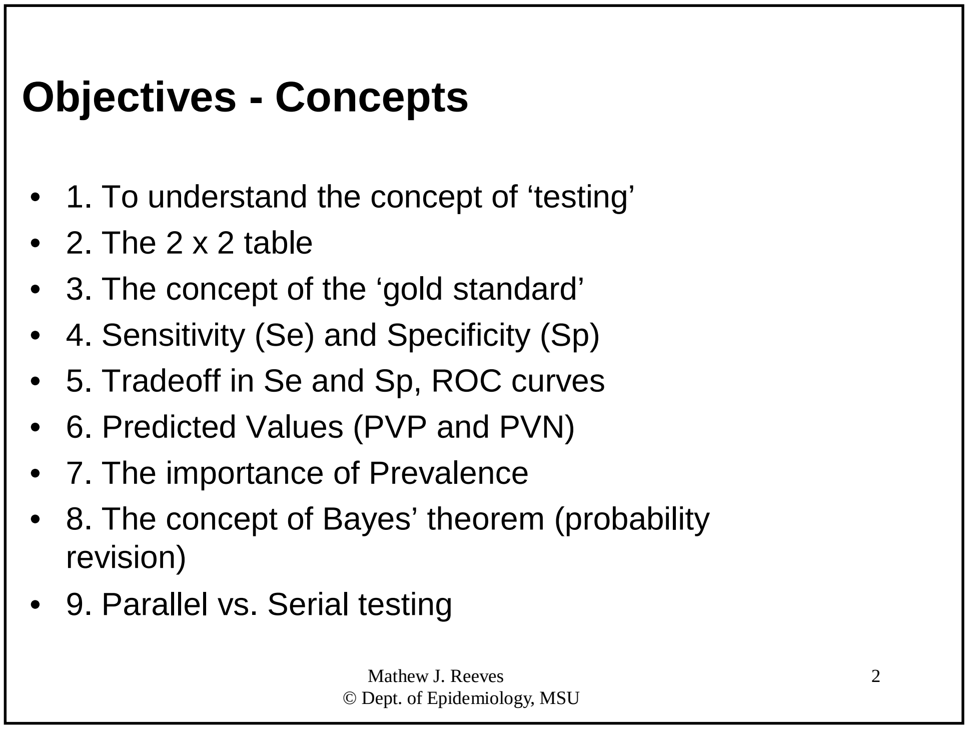


EPI-546 Block I

Lecture – Diagnosis/Clinical Testing

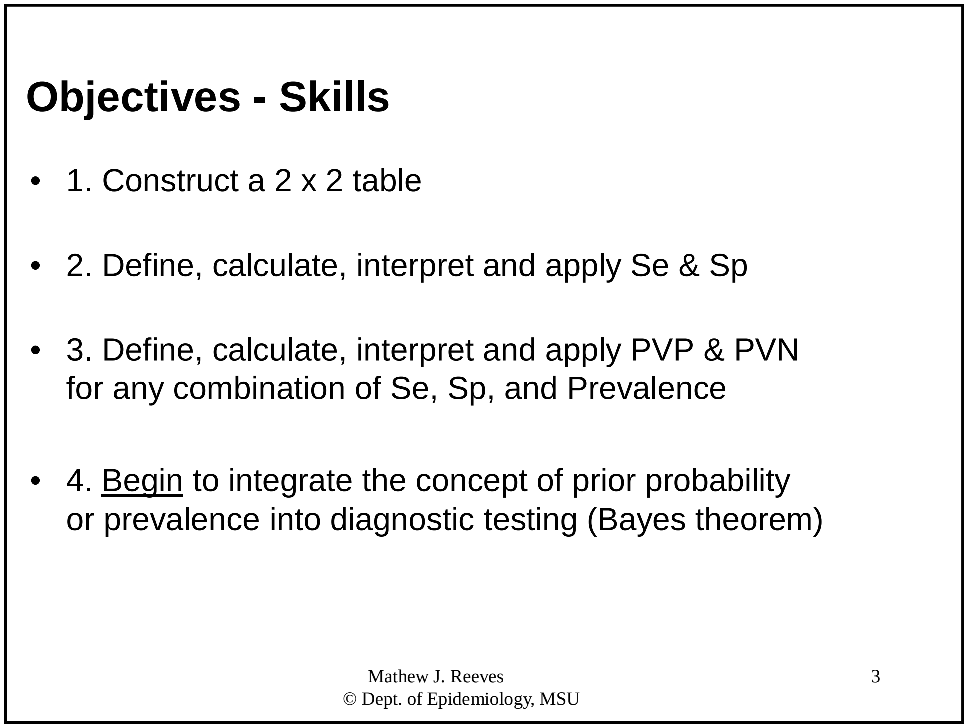
Understanding the process of diagnosis and clinical testing

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology



Objectives - Concepts

- 1. To understand the concept of ‘testing’
- 2. The 2 x 2 table
- 3. The concept of the ‘gold standard’
- 4. Sensitivity (Se) and Specificity (Sp)
- 5. Tradeoff in Se and Sp, ROC curves
- 6. Predicted Values (PVP and PVN)
- 7. The importance of Prevalence
- 8. The concept of Bayes’ theorem (probability revision)
- 9. Parallel vs. Serial testing

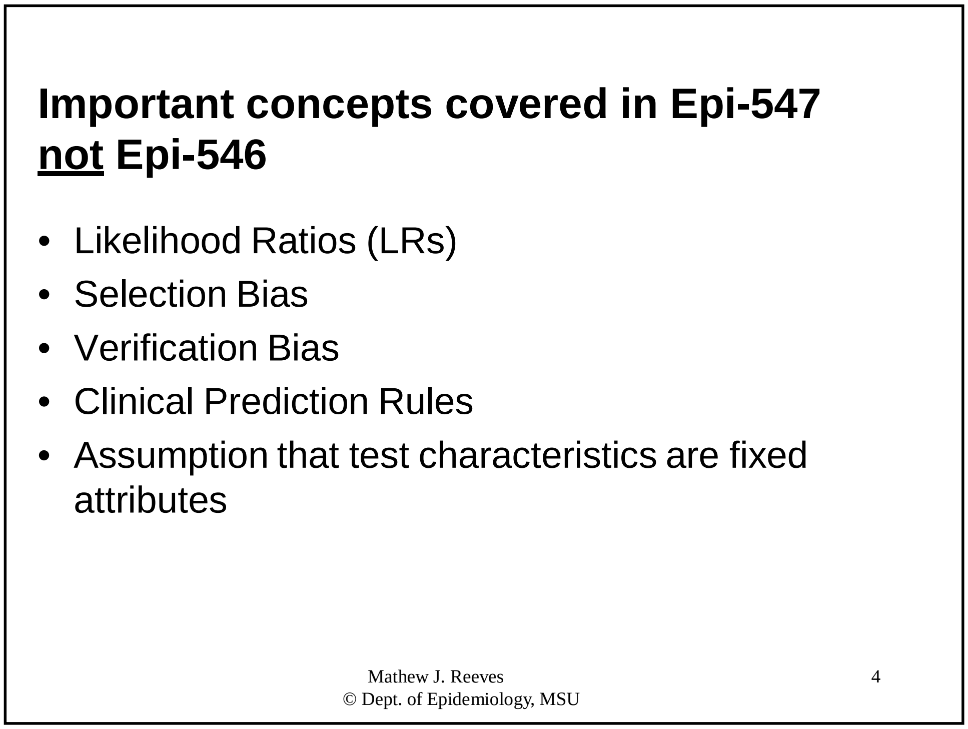


Objectives - Skills

- 1. Construct a 2 x 2 table
- 2. Define, calculate, interpret and apply Se & Sp
- 3. Define, calculate, interpret and apply PVP & PVN for any combination of Se, Sp, and Prevalence
- 4. <u>Begin</u> to integrate the concept of prior probability or prevalence into diagnostic testing (Bayes theorem)

Mathew J. Reeves
© Dept. of Epidemiology, MSU

3



Important concepts covered in Epi-547 <u>not</u> Epi-546

- Likelihood Ratios (LRs)
- Selection Bias
- Verification Bias
- Clinical Prediction Rules
- Assumption that test characteristics are fixed attributes

Mathew J. Reeves
© Dept. of Epidemiology, MSU

4

Riding a bike vs. mastering the balance beam……

Clinical Diagnosis and Clinical Testing

- *Diagnosis: the process of discovering a patient’s underlying disease status* by:
  - ascertaining the patient’s history, signs and symptoms, choosing appropriate tests, interpreting the results, and making correct conclusions.
  - Highly complicated, not well understood process.
- *Testing: the application of ‘clinical test information’ to infer disease status:*
  - ‘Clinical test information’ refers to <u>any</u> piece of information not just laboratory or diagnostic tests!!

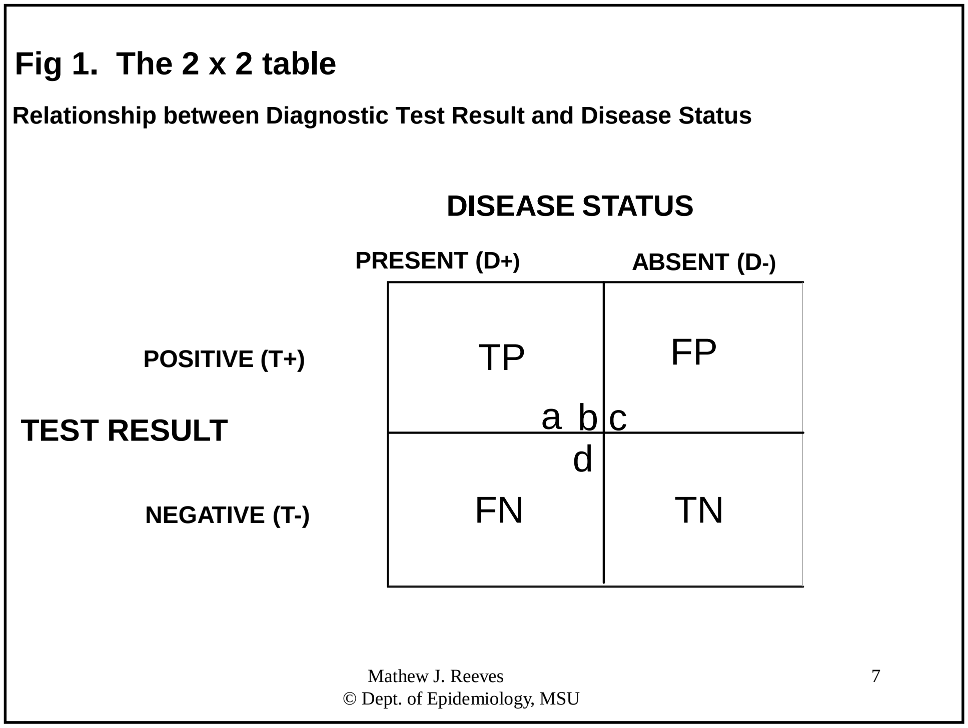
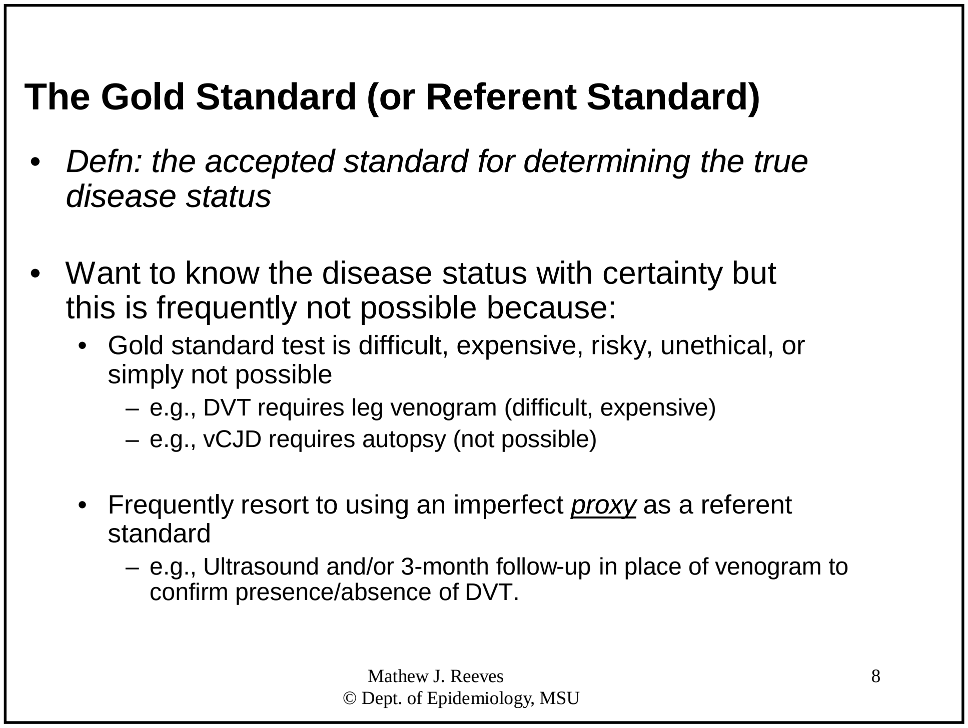


Fig 1. The 2 x 2 table

Relationship between Diagnostic Test Result and Disease Status

		DISEASE STATUS	
		PRESENT (D+)	**ABSENT (D-)**
TEST RESULT	**POSITIVE (T+)**	TP	FP
	NEGATIVE (T-)	FN	TN

a b c
d



The Gold Standard (or Referent Standard)

- *Defn: the accepted standard for determining the true disease status*

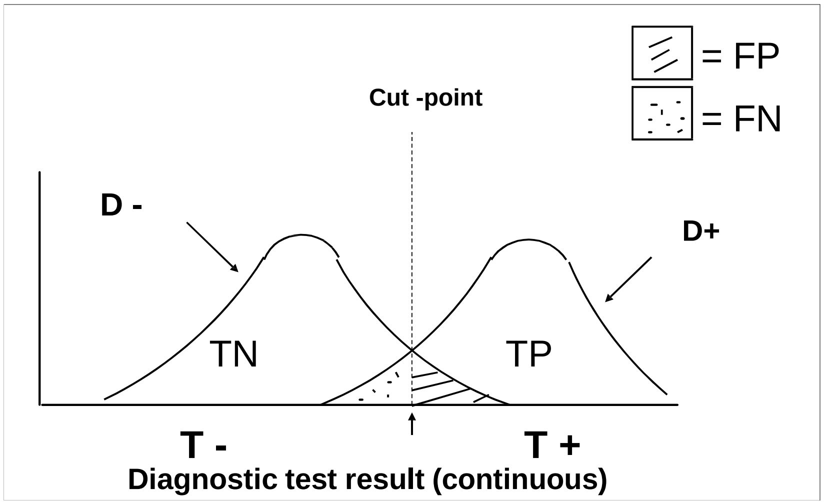
- Want to know the disease status with certainty but this is frequently not possible because:
  - Gold standard test is difficult, expensive, risky, unethical, or simply not possible
    - e.g., DVT requires leg venogram (difficult, expensive)
    - e.g., vCJD requires autopsy (not possible)
  - Frequently resort to using an imperfect *proxy* as a referent standard
    - e.g., Ultrasound and/or 3-month follow-up in place of venogram to confirm presence/absence of DVT.

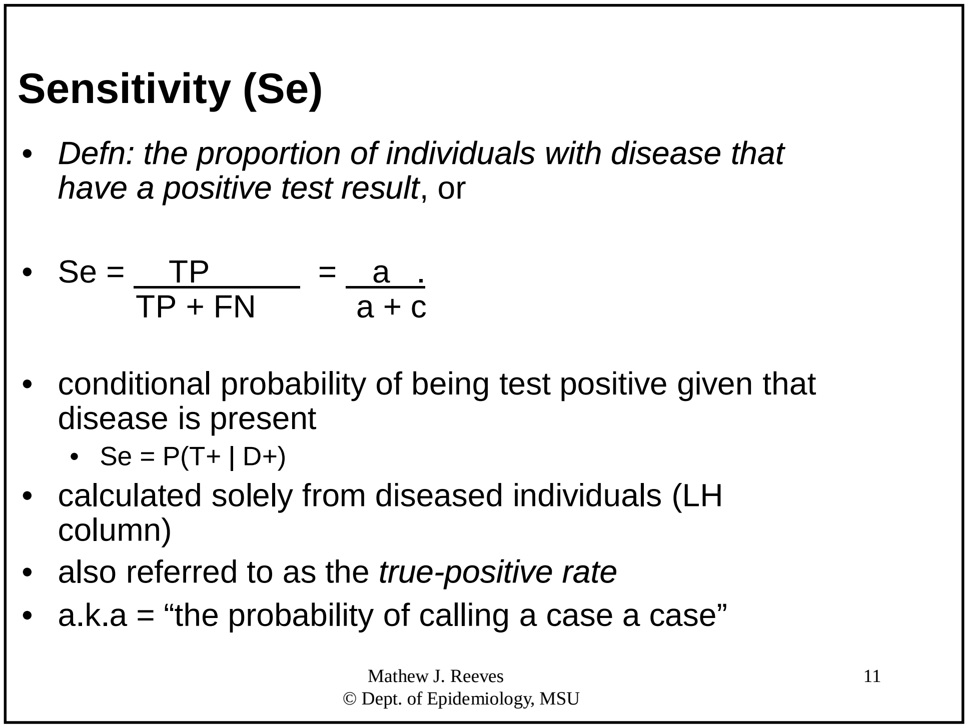
Sensitivity (Se) & Specificity (Sp)

- Interpretation of diagnostic tests is concerned with comparing the relative frequencies and “costs” of the incorrect results (FNs and FPs) versus the correct results (TPs and TNs).

- Degree of overlap is a measure of the test’s <u>effectiveness</u> or <u>discriminating ability</u> which is quantified by Se and Sp.

Fig 2. <u>Results for a Typical Diagnostic Test Illustrating Overlap Between Disease (D+) and Non-disease (D-) Populations</u>





Sensitivity (Se)

- *Defn: the proportion of individuals with disease that have a positive test result*, or
- Se = $\dfrac{TP}{TP + FN} = \dfrac{a}{a + c}$.
- conditional probability of being test positive given that disease is present
    - Se = P(T+ | D+)
- calculated solely from diseased individuals (LH column)
- also referred to as the *true-positive rate*
- a.k.a = "the probability of calling a case a case"

Mathew J. Reeves
© Dept. of Epidemiology, MSU

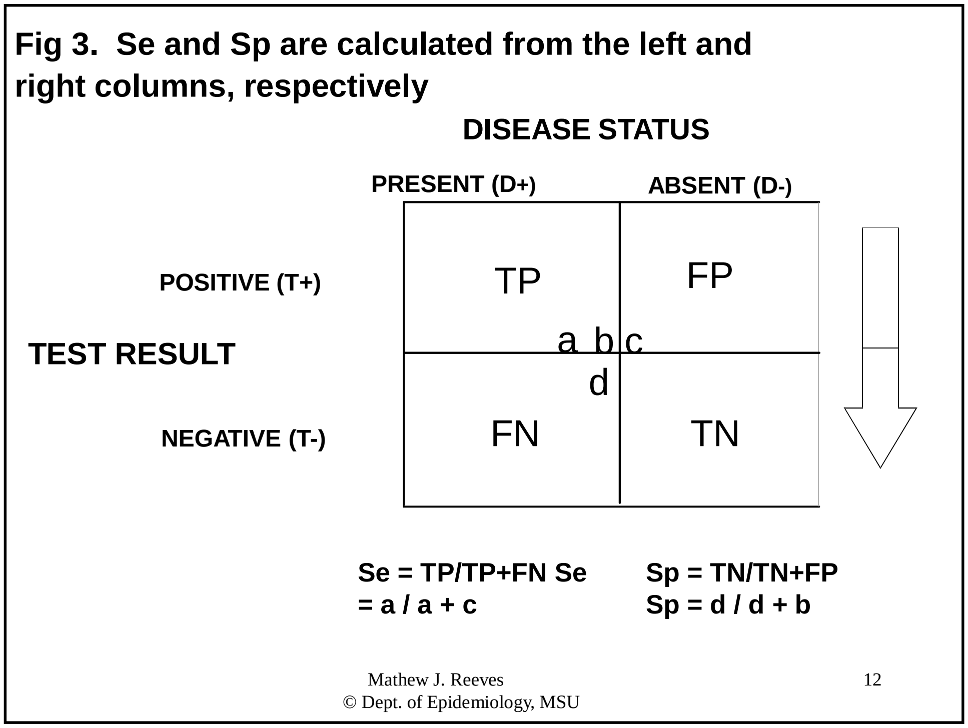


Fig 3. Se and Sp are calculated from the left and right columns, respectively

DISEASE STATUS

		PRESENT (D+)	**ABSENT (D-)**
TEST RESULT	**POSITIVE (T+)**	TP	FP
	NEGATIVE (T-)	FN	TN

a b c
d

Se = TP/TP+FN Se = a / a + c

Sp = TN/TN+FP Sp = d / d + b

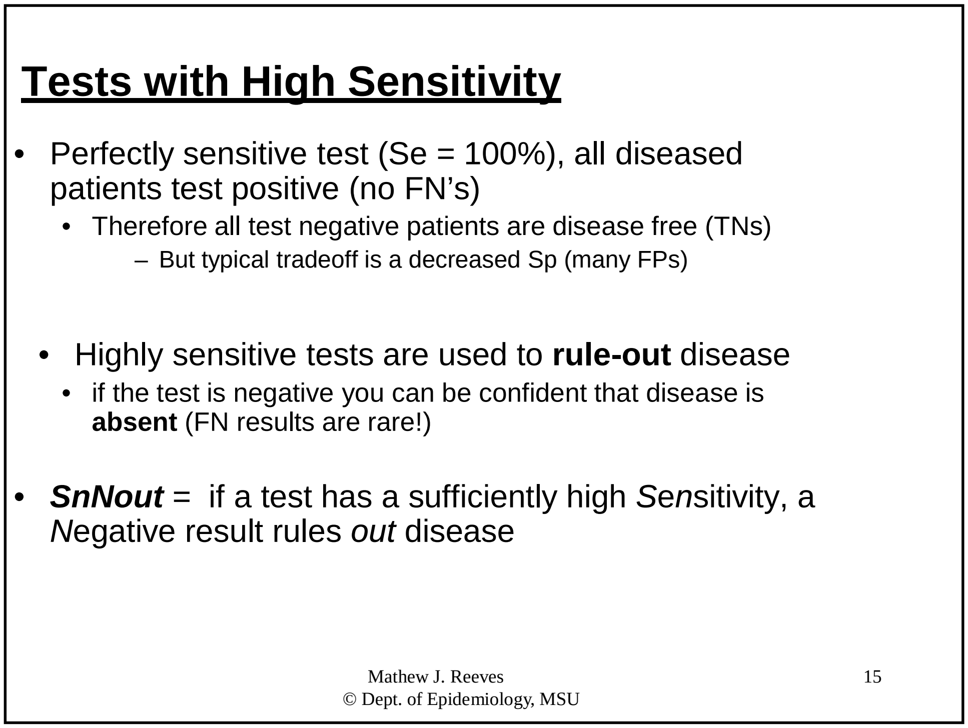
Mathew J. Reeves
© Dept. of Epidemiology, MSU

Specificity (Sp)

- *Defn: the proportion of individuals without disease that have a negative test result*, or
- $Sp = \frac{TN}{TN + FP} = \frac{d\ .}{d + b}$
- conditional probability of being test negative given that disease is absent
  - Sp = P(T-|D-)
- calculated solely from non-diseased individuals (RH Column).
- also referred to as the *true-negative rate*.
- a.k.a = "the probability of calling a control a control"

Fig 4. Se & Sp of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995)

		DVT		
		PRESENT (D+)	**ABSENT (D-)**	
D- dimer test	**POS (T+)**	47	37	84
		a b c d		
	NEG (T-)	6	124	130
		Se = 47/53 = 89%	Sp = 124/161 = 77%	214



Tests with High Sensitivity

- Perfectly sensitive test (Se = 100%), all diseased patients test positive (no FN’s)
    - Therefore all test negative patients are disease free (TNs)
        - – But typical tradeoff is a decreased Sp (many FPs)

- Highly sensitive tests are used to **rule-out** disease
    - if the test is negative you can be confident that disease is **absent** (FN results are rare!)

- ***SnNout*** = if a test has a sufficiently high *S*e*n*sitivity, a *N*egative result rules *out* disease

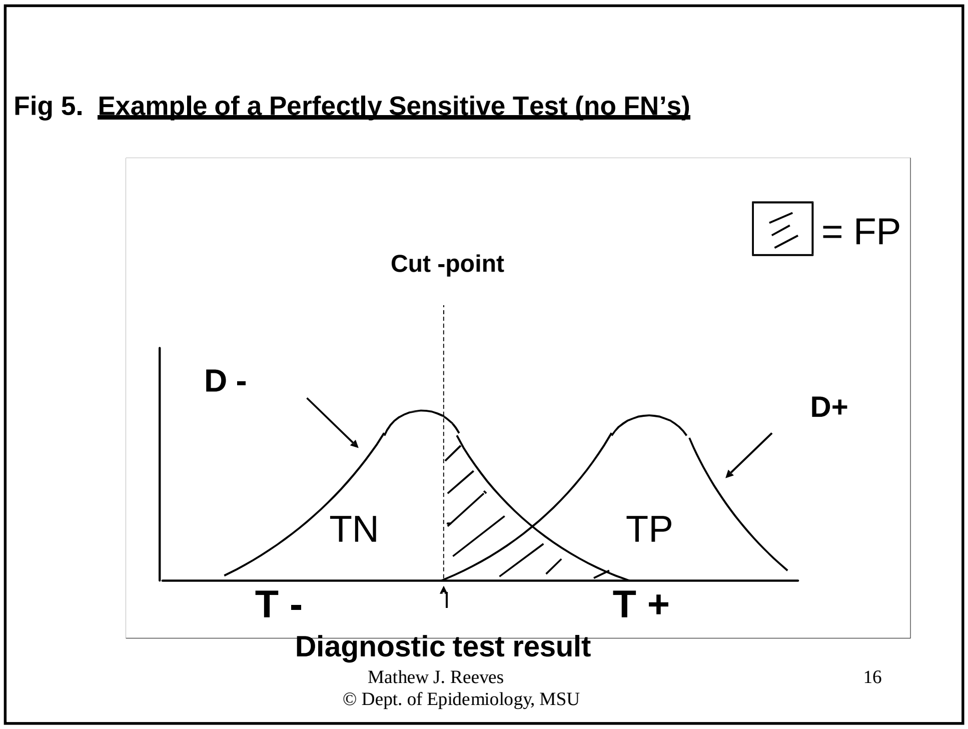
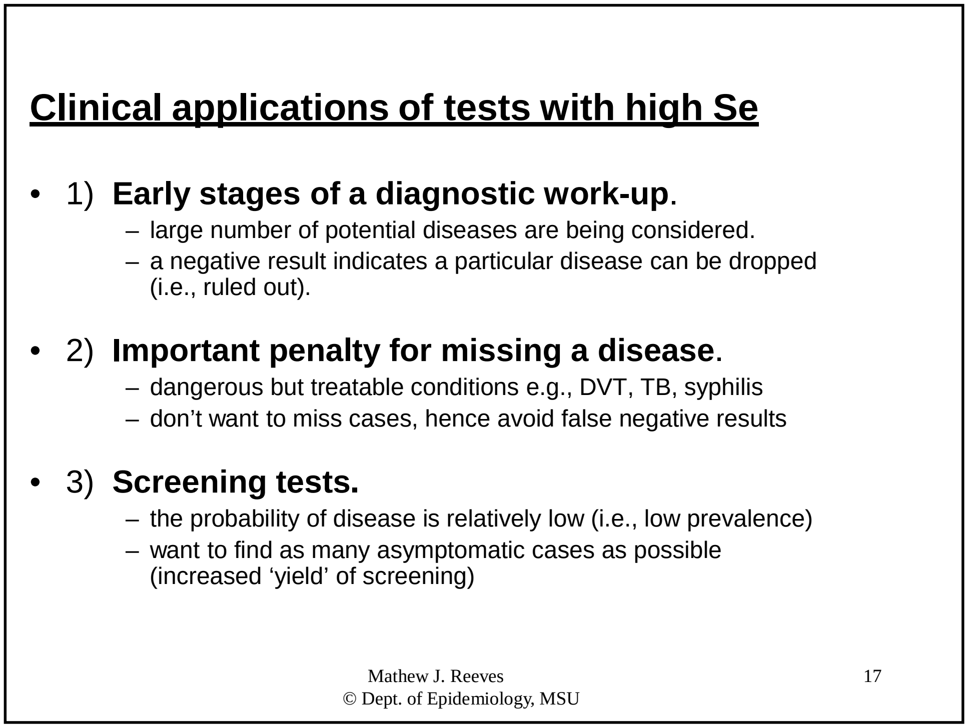


Fig 5. Example of a Perfectly Sensitive Test (no FN’s)



Clinical applications of tests with high Se

- 1) **Early stages of a diagnostic work-up**.
  - large number of potential diseases are being considered.
  - a negative result indicates a particular disease can be dropped (i.e., ruled out).
- 2) **Important penalty for missing a disease**.
  - dangerous but treatable conditions e.g., DVT, TB, syphilis
  - don’t want to miss cases, hence avoid false negative results
- 3) **Screening tests.**
  - the probability of disease is relatively low (i.e., low prevalence)
  - want to find as many asymptomatic cases as possible (increased ‘yield’ of screening)

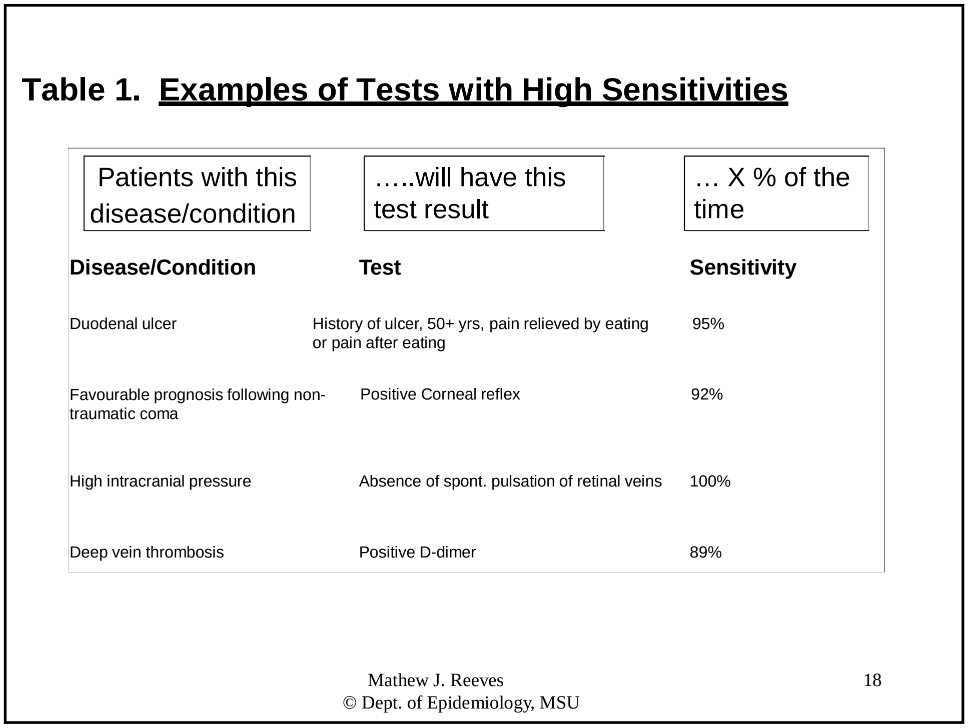


Table 1. Examples of Tests with High Sensitivities

Patients with this disease/condition	…..will have this test result	… X % of the time
Disease/Condition	**Test**	**Sensitivity**
Duodenal ulcer	History of ulcer, 50+ yrs, pain relieved by eating or pain after eating	95%
Favourable prognosis following non-traumatic coma	Positive Corneal reflex	92%
High intracranial pressure	Absence of spont. pulsation of retinal veins	100%
Deep vein thrombosis	Positive D-dimer	89%

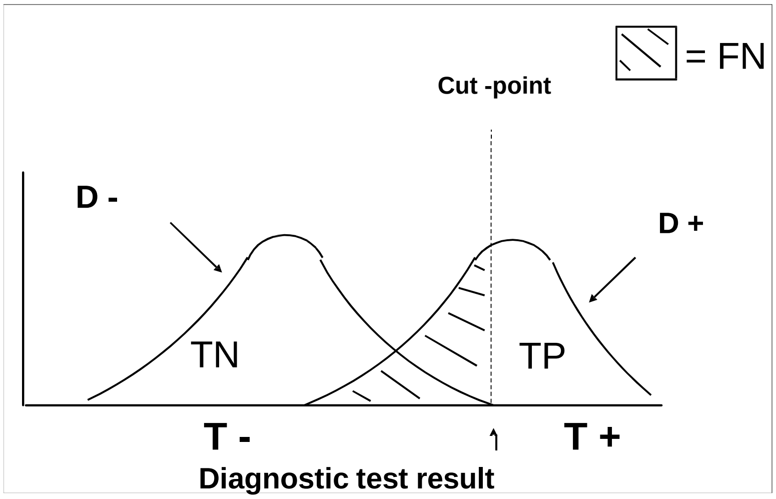
Tests with High Specificity

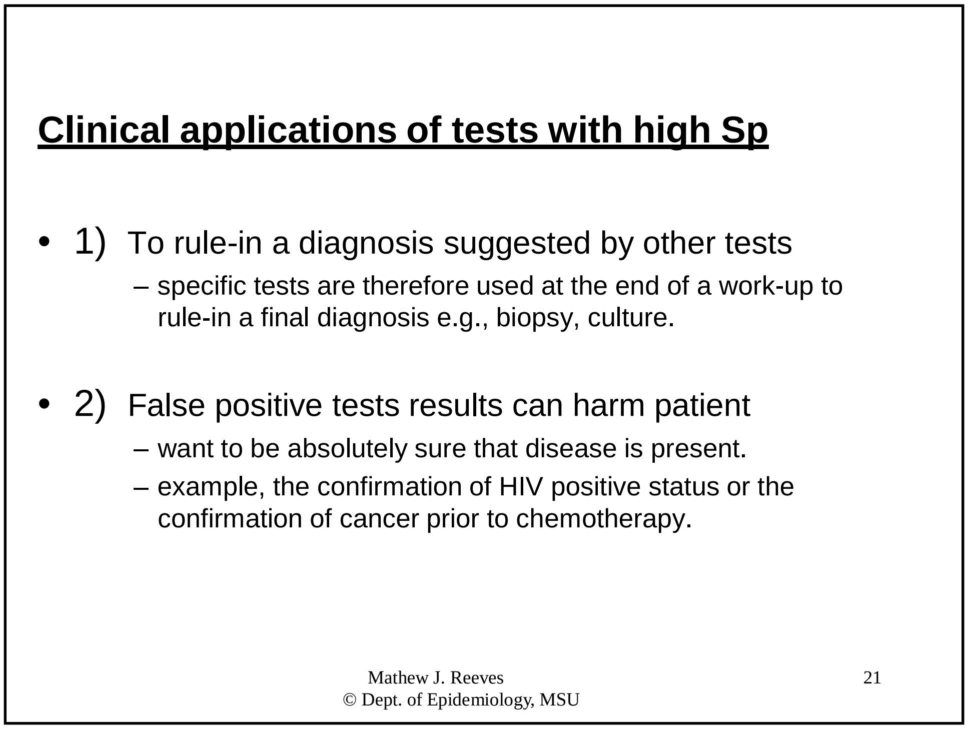
- Perfectly specific test (Sp = 100%), all non-diseased patients test negative (no FP’s)
  - Therefore all test positive patients have disease (TPs)
  - But typical tradeoff is large number of FNs

- Highly specific tests are used to **rule-in** disease
  - if the test is positive you can be confident that disease is **present** (FPs are rare).

- ***SpPin*** = if a test has a sufficiently high *Sp*ecificity, a *P*ositive result rules *in* disease.

Fig 6. Example of a Perfectly Specific Test (No FP’s)





Clinical applications of tests with high Sp

- 1) To rule-in a diagnosis suggested by other tests
  - – specific tests are therefore used at the end of a work-up to rule-in a final diagnosis e.g., biopsy, culture.
- 2) False positive tests results can harm patient
  - – want to be absolutely sure that disease is present.
  - – example, the confirmation of HIV positive status or the confirmation of cancer prior to chemotherapy.

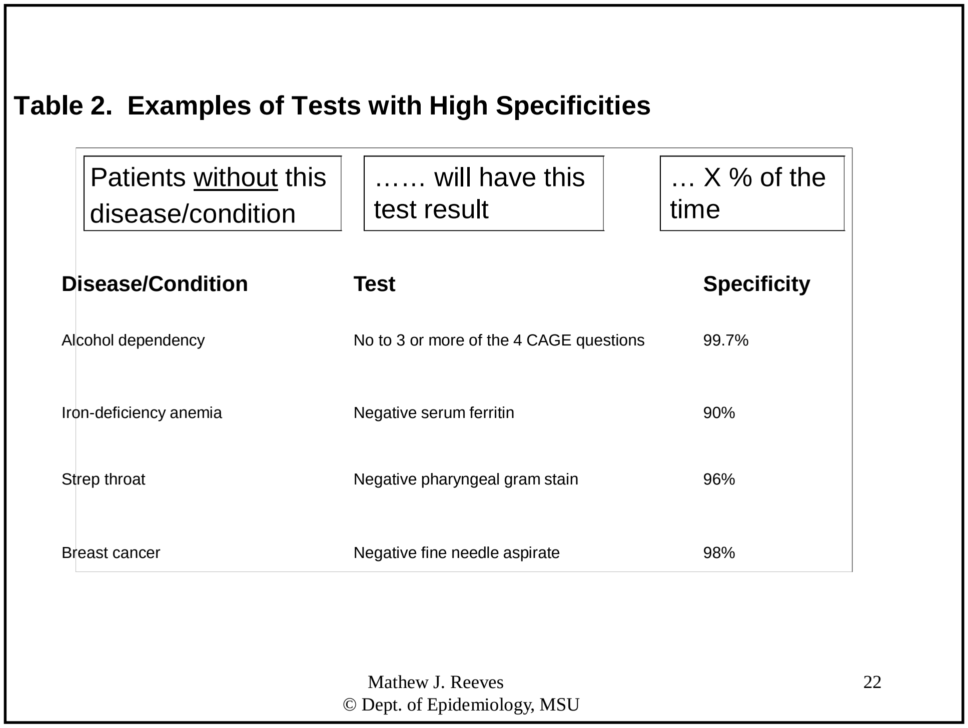


Table 2. Examples of Tests with High Specificities

Patients <u>without</u> this disease/condition	…… will have this test result	… X % of the time
Disease/Condition	**Test**	**Specificity**
Alcohol dependency	No to 3 or more of the 4 CAGE questions	99.7%
Iron-deficiency anemia	Negative serum ferritin	90%
Strep throat	Negative pharyngeal gram stain	96%
Breast cancer	Negative fine needle aspirate	98%

Trade off between Se and Sp

- Obviously we'd like tests with both high Se and Sp (> 95%), but this is rarely possible
- An inherent trade-off exists between Se and Sp (if you increase one the other must decrease)
- Whenever clinical data take on a range of values the location of the cut-point is arbitrary
  - Location should depend on the purpose of the test
    - Methods exist to calculate the best cut-point based on the frequency and relative "costs" of the FN and FP results
  - Trade-off between Se and Sp is demonstrated on ROC curve (refer to course notes and FF text)

Fig 7. A ROC Curve (2-hr post-prandial blood sugar for the diagnosis of Diabetes)

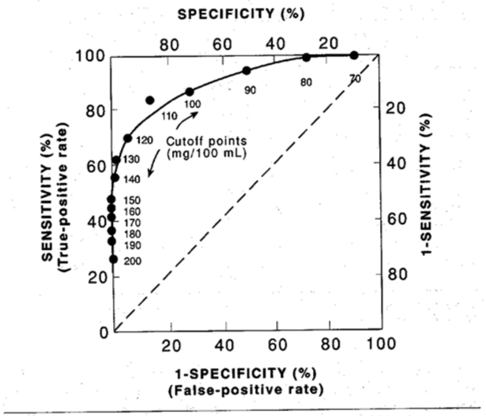


Figure 3.4. A ROC curve. The accuracy of 2-hr postprandial blood sugar as a diagnostic test for diabetes mellitus. (Data from Public Health Service. Diabetes program guide. Publication no. 506. Washington, DC: U.S. Government Printing Office, 1960.)

Predictive Values

- In terms of conditional probabilities Se and Sp are defined as:
  - Se = P(T+|D+)   Sp = P(T-|D-)
- **Problem**: can only be calculated if the true disease status is known!
- **Bu**t: the clinician is using a test precisely because the disease status is unknown!
- So, clinician actually wants **the conditional probability of disease given the test result**, OR
  - P(D+|T+) and P(D-|T-)

Predictive Value Positive (PVP)

- *Defn: the probability of disease in a patient with a positive (abnormal) test*.
- PVP = $\frac{TP}{TP + FP} = \frac{a}{a + b}$
- calculated solely from test positive individuals (top row of 2 x 2 table)
- conditional probability of being diseased given the test was positive, or PVP = P(D+|T+)
- N.B. the link between Sp and PVP via the FP rate (cell b). A highly specific test **rules-in** disease (think *SpPin*) because PVP is maximized.

Fig 8. PVP and PVN are calculated from the top and bottom rows, respectively

DISEASE STATUS

		PRESENT (D+)	**ABSENT (D-)**	
TEST RESULT	**POSITIVE (T+)**	TP a b	FP c	**PVP = TP/TP+FP** **PVP = a / a + b**
	NEGATIVE (T-)	d FN	TN	**PVN = TN/TN+FN** **PVN = d / d + c**

Predictive Value Negative (PVN)

- *Defn: the probability of not having disease when the test result is negative (normal).*

- PVN = $\frac{TN}{TN + FN} = \frac{d}{d + c}$.

- calculated solely from test negative individuals (bottom row of 2 x 2 table)

- conditional probability of not being diseased given the test was negative or PVN = P(D-|T-)

- clinically we are also interested in the complement of the PVN i.e., 1- PVN, which is the probability of having disease despite testing negative
  - 1- PVN = P(D+|T-)

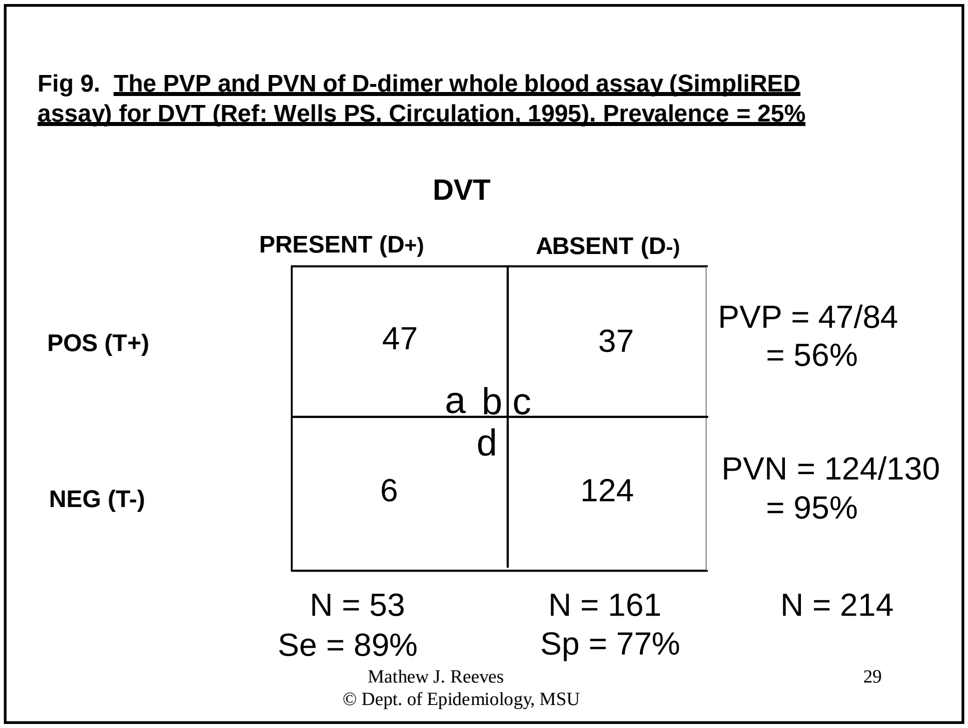


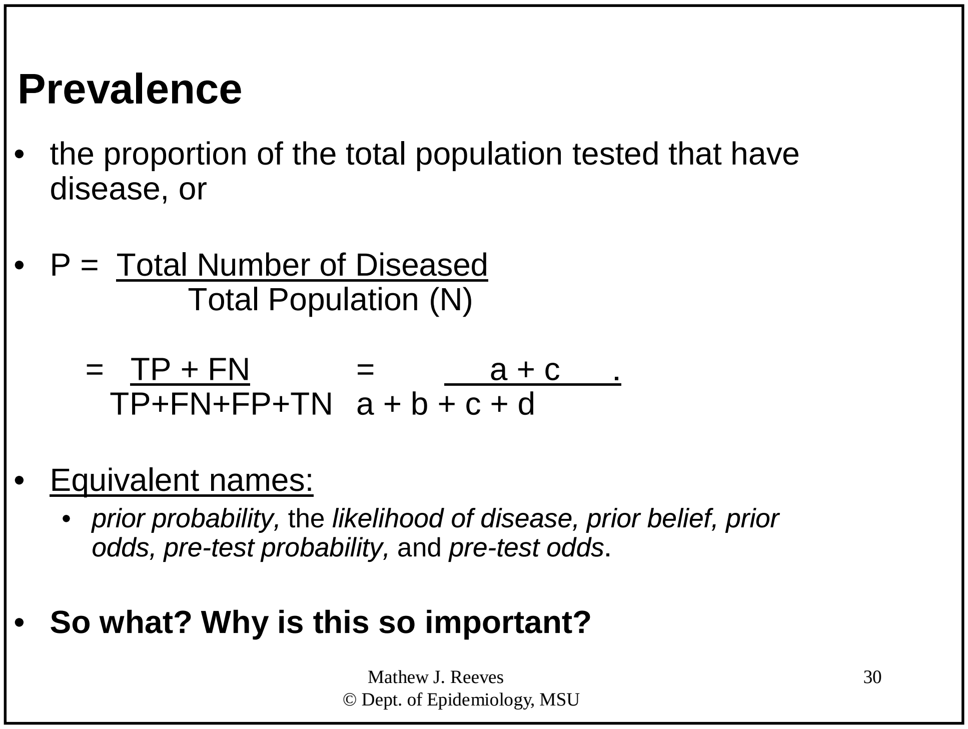
Fig 9. The PVP and PVN of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995). Prevalence = 25%

DVT

	PRESENT (D+)	ABSENT (D-)	
POS (T+)	47	37	PVP = 47/84 = 56%
NEG (T-)	6	124	PVN = 124/130 = 95%

	N = 53
	Se = 89%
		N = 161
		Sp = 77%
			N = 214

a b c
d



Prevalence

- the proportion of the total population tested that have disease, or
- P = $\dfrac{\text{Total Number of Diseased}}{\text{Total Population (N)}}$

$$= \frac{TP + FN}{TP+FN+FP+TN} = \frac{a + c}{a + b + c + d}$$

- Equivalent names:
  - *prior probability,* the *likelihood of disease, prior belief, prior odds, pre-test probability,* and *pre-test odds*.
- **So what? Why is this so important?**

Fig 10. The PVP and PVN of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995). Prevalence = 5%

		DVT		
		PRESENT (D+)	**ABSENT (D-)**	
D-dimer	**POS (T+)**	10	47	PVP = 10/57 = 18%
	NEG (T-)	1	156	PVN = 156/157 = 99.4%

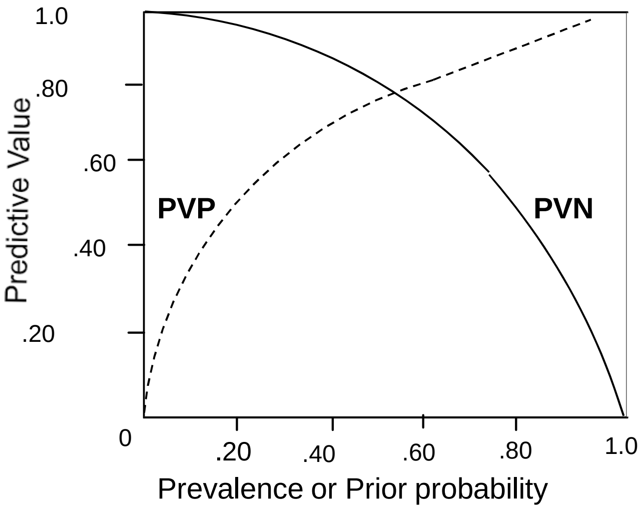
		N = 11
		Se = 89%
			N = 203
			Sp = 77%
				N = 214

a b c
d

Importance of Prevalence or Prior Probability

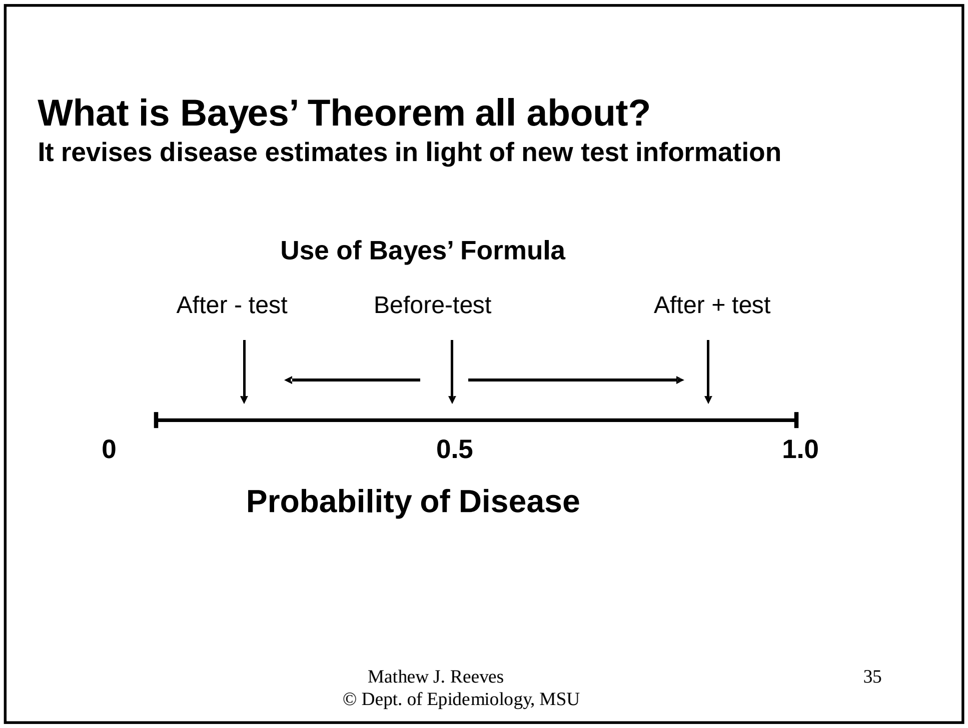
- Has a **dramatic** influence on predictive values
- Prevalence can <u>vary widely</u> from hospital to hospital, clinic to clinic, or patient to patient
- The same test (meaning the same Se and Sp) when applied under different scenarios (meaning different prevalence's) can give very different results (meaning different PVP and PVN!)
- N.B. Prior probability represents what the clinician *believes* (prior belief or clinical suspicion)
  - set by considering the practice environment, patients history, physical examination findings, experience and judgment etc

Fig 11. The PVP and PVN as a Function of Prevalence for a Typical Diagnostic Test



Bayes' Theorem

- Bayes' theorem = a unifying methodology for interpreting clinical test results.
- Bayes' formulae or equations are used to calculate PVP and PVN for any combination of Se, Sp, and Prev.
- PVP = $$\frac{[Se\ .\ Prev]}{[Se\ .\ Prev] + [(1 - Sp)\ .\ (1 - Prev)]}$$
- PVN = $$\frac{[Sp\ .\ (1 - Prev)]}{[Sp\ .\ (1 - Prev)] + [(1 - Se)\ .\ Prev]}$$



What is Bayes' Theorem all about?

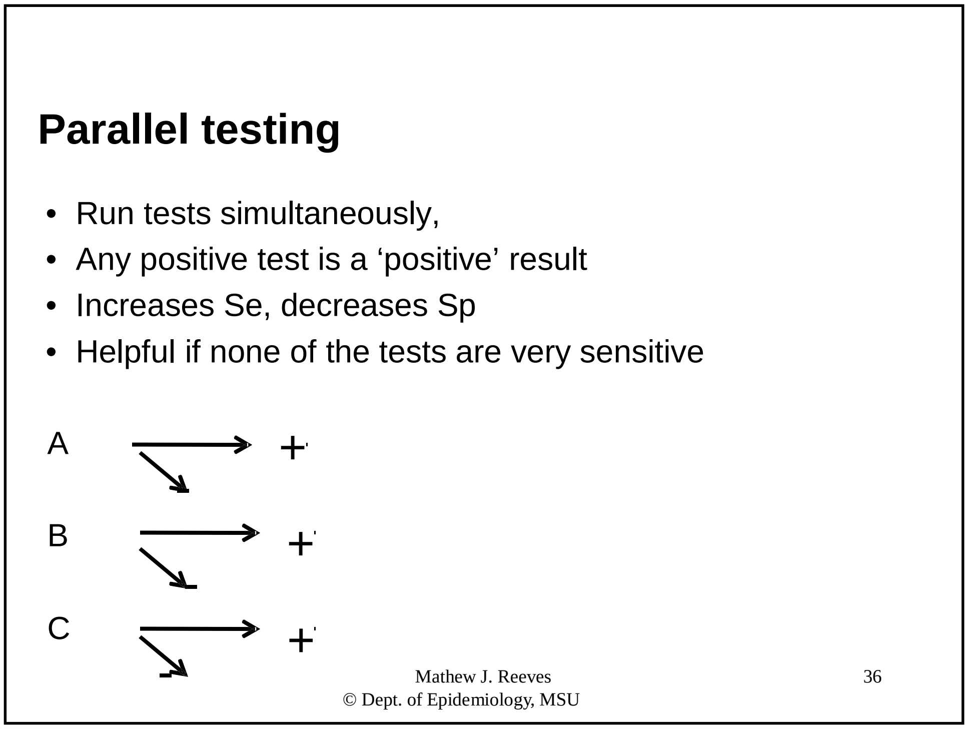
It revises disease estimates in light of new test information

Use of Bayes' Formula

After - test　　Before-test　　After + test

0　　**0.5**　　**1.0**

Probability of Disease



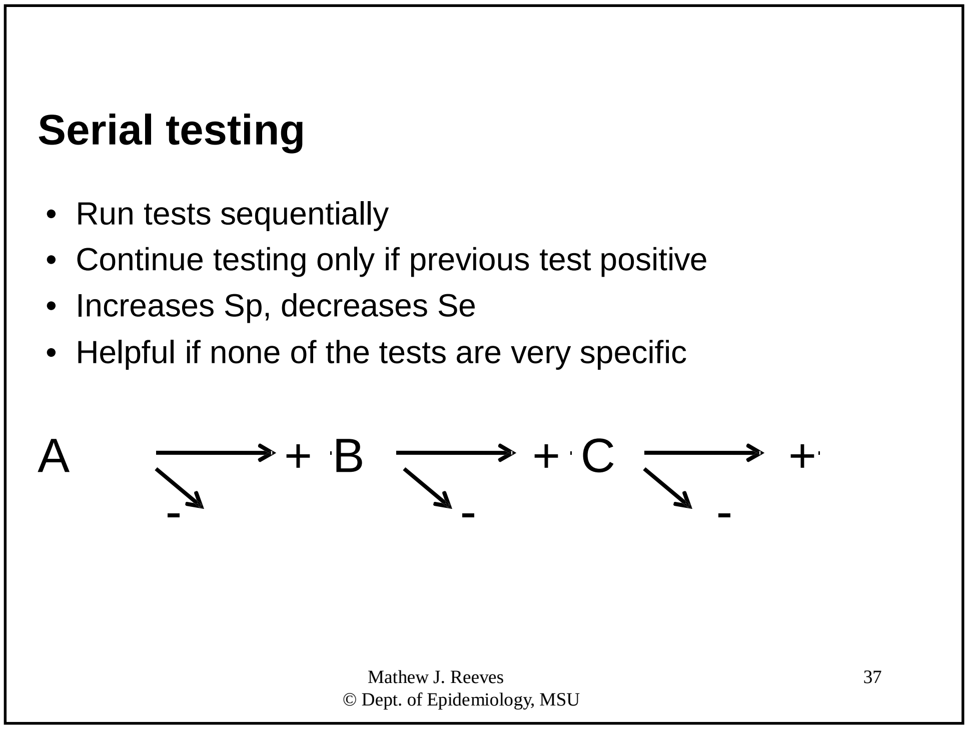
Parallel testing

- Run tests simultaneously,
- Any positive test is a 'positive' result
- Increases Se, decreases Sp
- Helpful if none of the tests are very sensitive

A → +
↘ –

B → +
↘ –

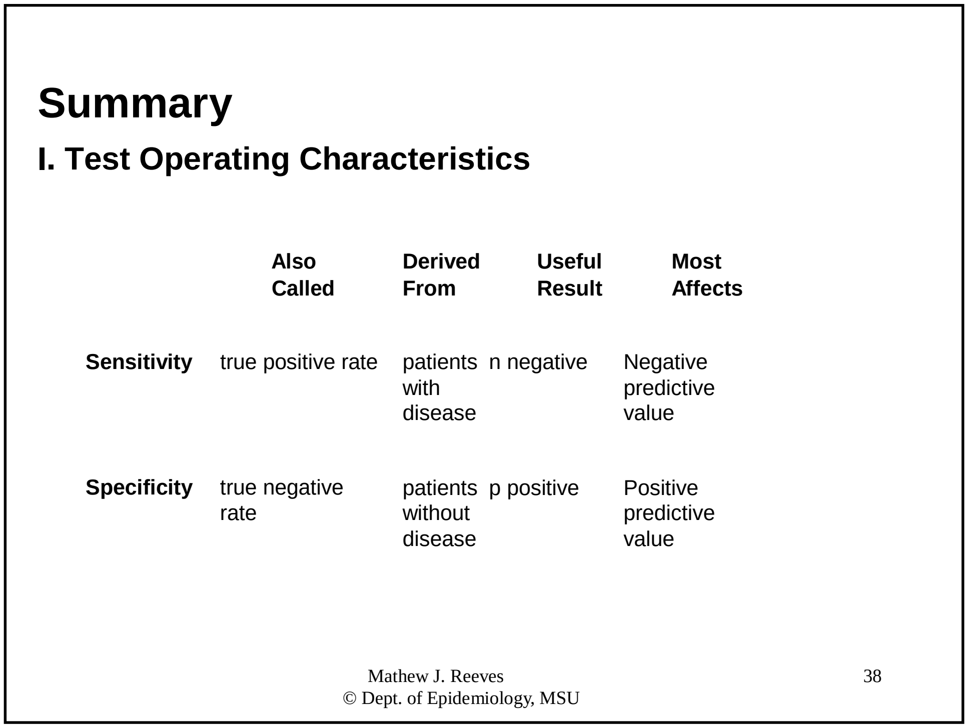
C → +
↘ –



Serial testing

- Run tests sequentially
- Continue testing only if previous test positive
- Increases Sp, decreases Se
- Helpful if none of the tests are very specific

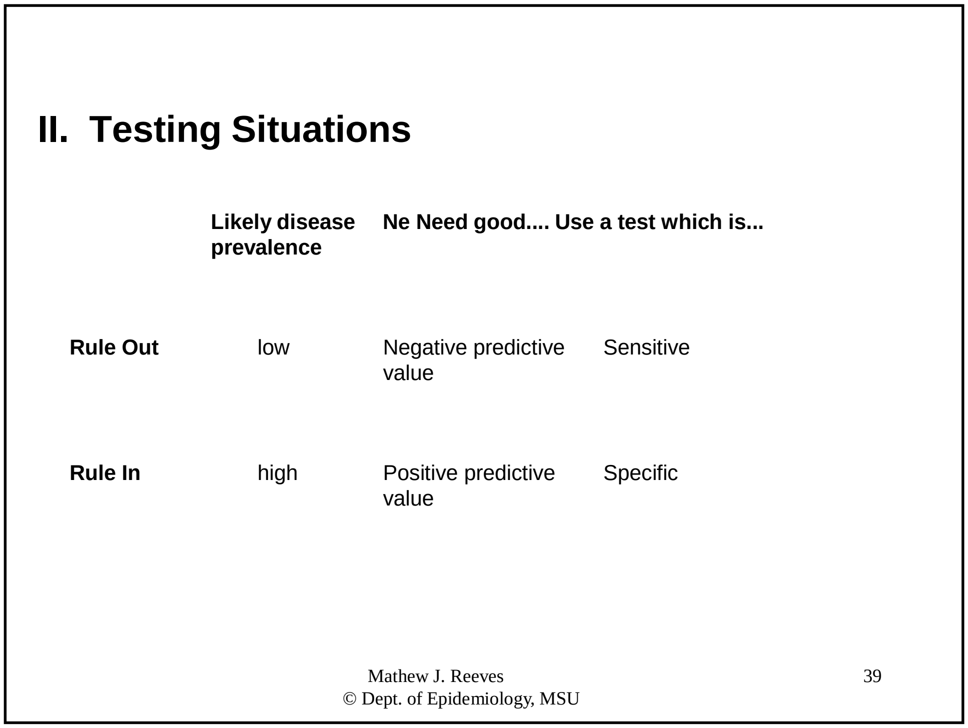
$$\begin{tikzcd} A \arrow[r] \arrow[dr] & + \; B \arrow[r] \arrow[dr] & + \; C \arrow[r] \arrow[dr] & + \\ & - & - & - \end{tikzcd}$$



Summary

I. Test Operating Characteristics

	Also Called	**Derived From**	**Useful Result**	**Most Affects**
Sensitivity	true positive rate	patients with disease	n negative	Negative predictive value
Specificity	true negative rate	patients without disease	p positive	Positive predictive value



II. Testing Situations

	Likely disease prevalence	Ne Need good....	Use a test which is...
Rule Out	low	Negative predictive value	Sensitive
Rule In	high	Positive predictive value	Specific

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes - Clinical Testing

Mat Reeves BVSc, PhD

Objectives

1. To understand the concept of 'testing'
2. Construct a 2 x 2 table for a diagnostic test
3. Understand the concept of the 'gold' (reference) standard
4. To define, calculate, interpret and apply sensitivity (Se) and specificity (Sp)
5. To understand the inherent tradeoff in Se and Sp,
6. To understand the construction and interpretation of the ROC curve
7. To define, calculate, interpret and apply predicted values (PVP and PVN)
8. The understand the role and importance of Prevalence on predictive values
9. The concept of Bayes' theorem (probability revision)
10. To understand the application of parallel and serial testing

Outline:

I. Clinical Testing
II. Clinical Test Characteristics
  A. Sensitivity (Se) and Specificity (Sp)
  B. Example Clinical Problem - Deep Vein Thrombosis
  C. Trade-off between Se and Sp
  D. ROC curves
III. Prevalence and Predictive Values
IV. Using Bayes' theorem to calculate predictive values
V. Multiple testing strategies

I. Clinical Testing (Diagnostic Strategies)

Diagnosis is the process of discovering a patient's underlying disease by ascertaining the patient's signs and symptoms, choosing appropriate tests, interpreting the results and arriving at (hopefully correct) conclusions. This is often a highly complicated process and the manner in which experienced clinicians arrive at a diagnosis is not well understood.

Hypothetico-deductive reasoning refers to the diagnostic strategy that nearly all clinicians use most of the time. It is defined as the formulation from the earliest clues, of a short list of potential diagnoses or actions, followed by the performance of those clinical and laboratory tests that best reduce the length of the list to come up with a final diagnosis. The list of possibilities is reduced by considering the evidence for and against each, discarding those which are very unlikely and conducting further tests to increase the likelihood of the most plausible candidates. The process as used by experienced
clinicians can be described in the following steps:

1. Formulate explanations (hypotheses) for the patient=s primary problem.
2. First consider those explanations that are *most* likely and/or those that are particularly harmful to miss (e.g., cerebral aneurysm for headache).
3. Simultaneously rule-out those that would be particularly harmful or catastrophic and try to rule-in those that are considered to be most likely.
4. Continue until the candidate list of explanations is shortened (i.e., 2-3) and/or one candidate disease is identified as having a very high likelihood (i.e., > 90% sure).

II. Clinical Test Characteristics

A. Sensitivity and Specificity

In clinical testing parlance a diagnostic test can be applied to any piece of clinical information whether obtained from the patient's history, the physical examination or use of diagnostic procedures such as radiography or electrocardiography.
To aid our discussion we will assume, initially, that both the disease and the diagnostic test have only two levels. So, the disease is either present (D+) or absent (D-) and the test is either positive (T+) (i.e., the test indicates that disease is present) or negative (T-) (i.e., the test indicates that the disease is absent).

There are 4 possible interpretations of these test results, two of which are correct - true positive (TP) and true negative (TN) and two of which are incorrect - false positive (FP) and false negative (FN). The relationship between these 4 test results is typically shown in the form of a two-by-two table (Figure 1).

Figure 1. Relationship between a Dichotomous Test Result and Disease Status

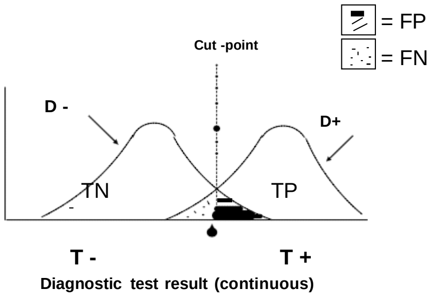
		DISEASE STATUS	
		PRESENT (D+)	ABSENT (D-)
TEST RESULT	POSITIVE (T+)	TP a b c	FP d
	NEGATIVE (T-)	FN	TN

Note that because of the inherent variability in biological systems there is no perfect test - false positive and false negative results always occur. The interpretation of diagnostic test results is essentially concerned with comparing the relative frequencies of the two incorrect results - the FNs and FPs to the two correct results - the TPs and TNs.

While some diagnostic tests results are inherently either positive or negative, for many common diagnostic tests the results are expressed on a continuous scale. In such cases, it is common to divide the continuum of values into either ‘abnormal’ or ‘normal’ findings, where ‘abnormal’ means a test measurement that is indicative of having disease, and ‘normal’ means indicative of not having disease. The point at which values are determined to be either normal or abnormal is called the cut-point. Tests will seldom be perfect in separating out diseased from non-diseased patients -

virtually all have some overlap between the two populations, as illustrated in Figure 2.

Figure 2. Results for a Typical Diagnostic Test Illustrating the Overlap between Diseased (D+) and Non-diseased (D-) Populations



To the extent that the two populations have similar measurements the test will not be able to discriminate between them. Conversely, the less the extent of overlap between the diseased and non-diseased populations the greater the discriminatory power of the test. The degree of overlap is therefore a *measure of the test effectiveness* and it is this that both sensitivity (Se) and specificity (Sp) quantity.

When reading an article about a diagnostic test, the presence or absence of disease has to be determined using some other source of information known as the "gold standard". The gold standard may involve obtaining a culture or biopsy, performing an elaborate diagnostic procedure such as a CAT scan, confirming the presence or absence of disease at surgery or post-mortem or simply determining the response to treatment or the results of long-term follow up.

Sensitivity: Sensitivity (Se) is defined as the proportion of individuals with disease that have a positive test result, or

$$\text{Se} = \frac{\text{True Positives}}{\text{True Positives + False Negatives}} = \frac{\text{TP}}{\text{TP + FN}} = \frac{a}{a + c}$$

Se is the conditional probability of being test positive given that disease is present, or Se = P(T+ | D+). Se is also referred to as the *true-positive rate*. Note that Se is calculated solely from diseased individuals.

Specificity: Specificity (Sp) is defined as the proportion of individuals without disease that have a negative test result, or

$$\text{Sp} = \frac{\text{True Negatives}}{\text{True Negatives + False Positives}} = \frac{\text{TN}}{\text{TN + FP}} = \frac{d}{d + b}$$

Sp is the conditional probability of being test negative given that disease is absent, or

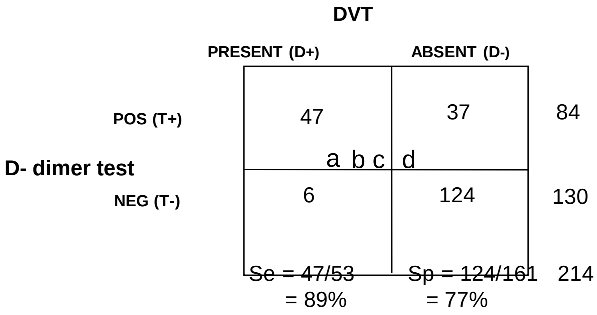
Sp = P(T-|D-). Sp is also referred to as the *true-negative rate*. Note that Sp is calculated solely from non-diseased individuals.

B. <u>Example Clinical Problem - Deep Vein Thrombosis</u>

Deep vein thrombosis (DVT) is a condition of active thrombosis in the deep venous system of one or both lower extremities that can lead to significant complications of pulmonary embolism, chronic venous insufficiency, and possibly death. The presence or absence of clinical signs and symptoms (pain, tenderness, swelling and edema) do not correlate well with the presence or absence of DVT. The gold standard diagnostic test is *ascending functional venography*, which at proficient centers, provides adequate visualization of the venous system in 95-98% of patients. The test is not perfect and can produce occasional false positive results. A negative or normal study however, virtually confirms that the patient is free of disease. There is a 2-4% risk of the procedure actually inducing DVT in patients who were originally free of the condition. Because of the technical requirements of venography, the fact that it is an imperfect, and that it is associated with some complications, other non-invasive techniques have been developed.

One such non-invasive test is the *D-dimer assay*. D-dimer is a specific degradation product of cross-linked fibrin and is therefore a marker of endogenous fibrinolysis (which would be expected to be elevated in the presence of a DVT). Several studies have shown that D-dimer assays have high Se but have only moderate Sp. The example data we will use (Figure 3) is from a Canadian study (Wells PS et al, Circulation 1995;91:2184-87) of 214 patients seen at two hospitals, all of whom had a whole blood assay for D-diner (SimpliRED) and underwent the gold standard test of contrast venography. The test characteristics were Se = 89% [95% CI = 77-96%], and Sp = 77% [95% CI = 63-80%].

Figure 3. <u>Se & Sp of D-dimer whole blood assay (SimpliRED) for DVT (Ref: Wells PS, Circulation, 1995)</u>



		DVT		
		PRESENT (D+)	ABSENT (D-)	
D- dimer test	POS (T+)	47	37	84
		a b c	d	
	NEG (T-)	6	124	130
		Se = 47/53 = 89%	Sp = 124/161 = 77%	214

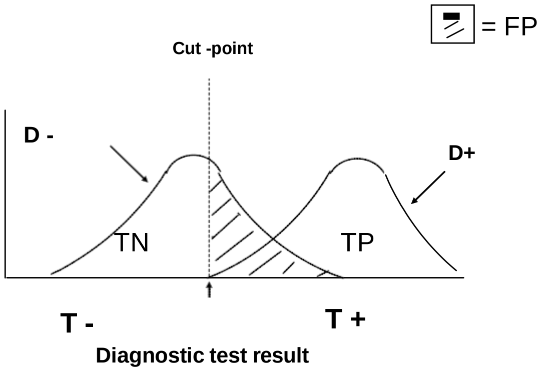
Clinicians need to take into account the inherent attributes of a test (i.e., its Se and Sp) before a test is selected and used. Since no test is perfect (i.e., 100% sensitive and 100% specific), the usefulness of a particular test will depend on what information the clinician wants to get from it.

Tests with High Sensitivity

For a perfectly sensitive test (i.e., Se = 100%), all diseased patients test positive - there are no false negative results (the false negative rate is zero) (Figure 4). Note that all test negative patients are disease free but that in this example, a sizeable proportion of the disease-free population test positive (= false positive).

A perfectly sensitive tests almost never exists, more typically we are interested in tests that have a high sensitivity (i.e., > 90%). Here, false negative results among diseased individuals are few in number - the vast majority of test negative patients are disease free. Highly sensitive tests are most useful to **rule-out** disease because if the test is negative you can be confident that disease is absent since FN results are rare.

Figure 4. Example of a Perfectly Sensitive Test (no FN's)



It is important to understand that a highly sensitive test **does not tell you if disease is present**, despite the fact that Se is calculated using only diseased individuals and that nearly all of these patients test positive! This is because Se provides no information regarding the number (or rate) of false positive results - this information is provided by the Sp of the test.

A highly sensitive test is therefore most helpful to a clinician when the test result is negative, because it rules out disease, whereas the interpretation of a positive result will depend on the rate of false positive results (see Specificity).

SnNout is a mnemonic designed to indicate that if a sign, symptom or other diagnostic tests has a sufficiently high *Sen*sitivity, a *N*egative result rules *out* disease.

There are three clinical scenarios when tests with high sensitivity should be selected:

1) in the early stages of a work-up when a large number of potential diseases are being considered. If the test has a high Se, a negative result indicates that a particular disease is very unlikely and it can therefore be dropped from consideration (i.e., ruled out).

2) when there is an important penalty for missing a disease. Examples include tuberculosis and syphilis, which are dangerous but treatable conditions. We would not want to miss these cases, hence we want a test that has a low number of false negative results (i.e., a highly Se test).

3) in screening tests where the probability of disease is relatively low (i.e., low disease prevalence) and the purpose is to discover asymptomatic cases of disease.

Table 1. Examples of Tests with High Sensitivities

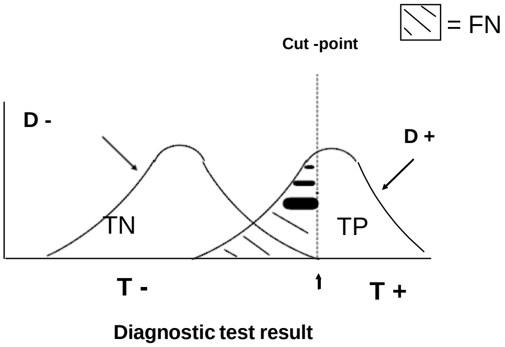
Patients with this disease/condition	**... will have this test result.....**	**... X % of the time**
Duodenal ulcer	History of ulcer, 50+ yrs, pain relieved by eating or pain after eating	95%
Favourable prognosis following non-traumatic coma	Positive Corneal reflex	92%
High intracranial pressure	Absence of spont. pulsation of retinal veins	100%
Deep vein thrombosis	Positive D-dimer	89%
Pancreatic cancer	Positive endoscopic retrograde cholangio-pancreatography (ERCP)	95%

Tests with High Specificity

For a perfectly specific test (i.e., Sp = 100%), all non-diseased patients test negative - there are no false positive results (the false positive rate is zero) (Figure 5). Also note that all test positive patients have disease but that in this example, a sizeable proportion of the diseased population test negative (= false negative).

Again, a perfectly specific tests almost never exists, more typically we are interested in tests that have a high specificity (i.e., > 90%). Here, false positive results among non-diseased individuals are few in number, so the vast majority of test positive patients have disease. Highly specific tests are most useful to **rule-in** disease, because if the test is positive you can be confident that disease is present since false positive results are rare.

Figure 5. Example of a Perfectly Specific Test (no FP's)



It is important to understand that a highly specific test **does not tell you if disease is absent**, despite the fact that Sp is calculated using only non-diseased individuals and that nearly all of them test negative. This is because Sp provides no information regarding the number (or rate) of false negative results - this information is provided by the Se of the test.

SpPin is a mnemonic designed to show that if a sign, symptom or other diagnostic tests has a sufficiently high *Sp*ecificity, a *P*ositive result rules *in* disease.

A highly specific test is therefore most helpful to a clinician when the test result is positive, since it rules-in disease. There are two clinical scenarios when tests with high specificity should be selected:

1) to rule-in a diagnosis that has been suggested by other tests - specific tests are therefore used at the end of a work-up to rule-in a final diagnosis e.g., biopsy, culture, CT scan.

2) when false positive tests results can harm the patient physically or emotionally. For example, the confirmation of HIV positive status or the confirmation of cancer prior to chemotherapy. A highly specific test is required when the clinician wants to be absolutely sure that a condition is present.

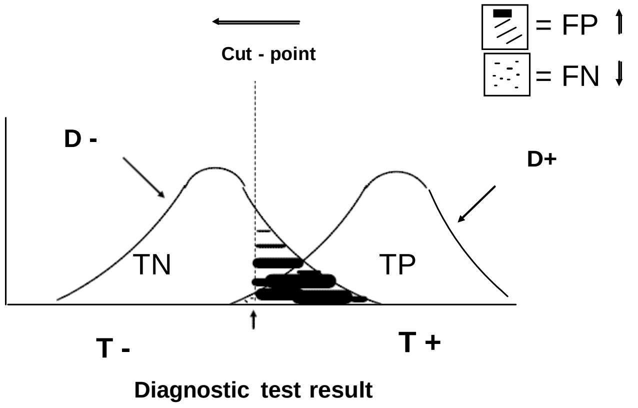
Table 2. Examples of Tests with High Specificities

Patients <u>without</u> this disease/condition	**... will have this test result.....**	**... X % of the time**
Alcohol dependency	No to 3 or more of the 4 CAGE questions	99.7%
Iron-deficiency anemia	Negative serum ferritin	90%
Breast cancer	Negative fine needle aspirate	98%
Strep throat	Negative pharyngeal gram stain	96%

C. Trade-off Between Sensitivity and Specificity

Because there is no such thing as a perfect test (a test that has no FP or FN results) there is an inherent trade-off between sensitivity and specificity. For clinical test results that have a continuous scale of measure, the location of the cut-point (the point on the continuum that divides normal and abnormal) is arbitrary and can be modified according to the purposes of the test. For example, in Figure 6 below, we can see that if the cut-point is made lower (i.e., moved to the left) there will be less FN results but more FP results - therefore Se will be increased at the expense of Sp. Figure 4 shows the extreme of this scenario, where the cut-point has been lowered to make Se = 100% at the expense of Sp. Conversely, Figure 5 shows the other extreme where the cut-point has been increased (moved to the right) to maximize Sp at the expense of Se. The trade-off between Sp and Se cannot be avoided and points to the fact that the ideal cut-point depends on what the purpose of the test is. A *Receiver Operator Characteristic (ROC) Curve* is a graphical way of illustrating the trade off between Se and Sp for various cut-points of a diagnostic test.

Fig 6. Trade-off: Lowering the Test Cut-point Increases Se but Decreases Sp

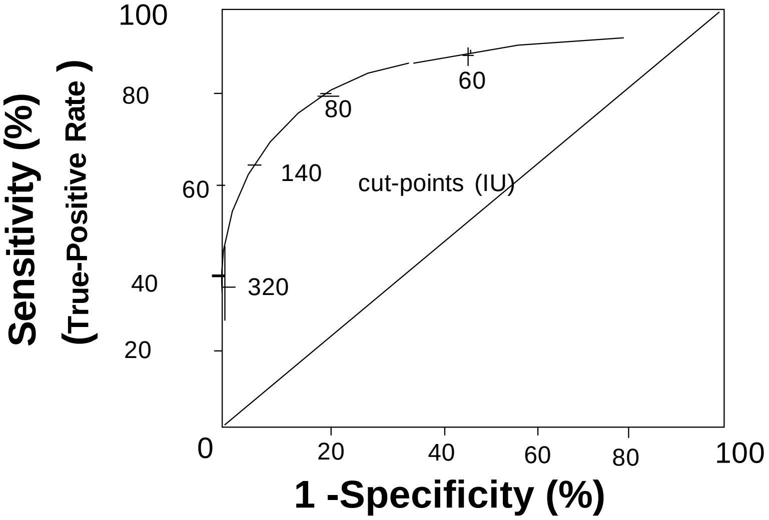


C. Receiver Operator Characteristic (ROC) Curves

There are two main uses of the ROC curve – to compare the accuracy of two or more tests, and to show the trade off between the Se and Sp as the cut-point is changed.

The ROC curve is constructed by plotting the sensitivity (or true positive rate) against the false positive rate (1 - Specificity), for a series of cut-points as illustrated in Figure 7 below which evaluates the utility of creatine kinase (CK) to diagnose acute myocardial infarction.

Figure 7. An Example Receiver Operator Characteristic Curve (The Accuracy of the CK Test in the Diagnosis of Myocardial Infarction)



The ROC curve is a good way of comparing the usefulness of different tests. The higher the sensitivity and specificity of a test the further the curve is pushed up into the top left hand corner of the box. Tests which discriminate best lie "further to the north-west", because they have both low FN rates (indicated by high Se) and low FP rates (indicated by a high Sp) (See Figure 3.5 in FF for an example of such a figure).

A test that has no discriminating ability has equal TP and FP rates which is indicated by the diagonal straight line in the above figure.[1] The ability of different tests to discriminate between diseased and non-diseased individuals can be quantified by calculating the Area Under the ROC Curve (AUROCC), which varies from 0.5 (no discriminating ability) to 1.0 (perfect accuracy).

[1] This is equivalent to a *likelihood ratio* (LR) of 1.0 (we will discuss LRs in Epi-547). Note that the slope of the ROC curve (i.e., the ratio of the TP Rate to the FP Rate) for any given cut-point is the likelihood ratio.

The ROC curve is also helpful in deciding on the best cut-point for a particular test. The choice of best cut-point is influenced by the likelihood of disease (i.e., its prevalence) and the relative costs (or risk-benefit ratio) associated with errors in diagnosis - both false positive and false negative. Advanced statistical decision theory can be applied to determine the optimal operating position on the ROC curve for a given test (the details of which are beyond this course), however, we can use the following intuition to understand the basic principle:

> If the cost of missing a diagnosis (a false negative result) is high compared to the cost of falsely labeling a healthy individual as diseased (a false positive result), then one would want to operate along the horizontal part of the curve (e.g., a cut-point of 60 CK units in Figure 7), since at this point FN results are minimized at the expense of FP results. A CK cut point of about 60 therefore maximizes sensitivity (i.e., ~90%) while providing reasonable specificity (i.e., ~50%)
>
> On the other hand, if the cost of falsely labeling a healthy person as diseased (a false positive result) is high compared to the cost of missing a diagnosis (a false negative result), then one would want to operate along the vertical part of the curve (e.g., a cut-point of 320 CK units in Figure 7), since at this point FP results are minimized. A CK cut point of 320 therefore maximizes specificity (i.e., ~99%) while providing moderate sensitivity (i.e., ~40%)

III. Prevalence and Predictive Values

While Sensitivity and Specificity are important concepts to understand they, unfortunately, don't tell the full story of clinical testing. In terms of conditional probabilities Se and Sp can be defined as:

$$Se = P(T+|D+)\ Sp$$

$$= P(T-|D-)$$

The problem is that these measures can only be calculated if the true disease status is known i.e., they are "conditional" on the disease status being either positive (for Se) or negative (for Sp). However, the clinician is using a test precisely because the true disease status of the patient is unknown! The clinician wants to know the conditional probability of disease given the test result e.g., P(D+|T+), hence Se and Sp are apparently not much use!

In order to use diagnostic test data to infer the true disease status of a patient, the clinician needs to understand the concepts of predictive values and prevalence:

Predictive Value Positive (PVP): Predictive value positive is the probability of disease in a patient with a positive (abnormal) test.

$$PVP = \frac{\text{True Positives}}{\text{True Positives + False Positives}} = \frac{TP}{TP + FP} = \frac{a}{a + b}$$

PVP is the conditional probability of being diseased given that the test was positive, or PVP = P(D+|T+). Note that Sp and PVP are "linked" in that they both provide information on the FP rate. A highly specific test helps to **rule-in** disease because PVP is maximized.

Predictive Value Negative (PVN): Predictive value negative is the probability of *not* having disease when the test result is negative (normal).

$$PVN = \frac{\text{True Negatives}}{\text{True Negatives + False Negatives}} = \frac{TN}{TN + FN} = \frac{d}{d + c}$$

PVN is the conditional probability of not being diseased given that the test was negative, or PVN = P(D-|T-). Note that Se and PVN are "linked" in that they both provide information on the FN rate. A highly sensitive test helps to **rule-out** disease because PVN is maximized.

From a clinical standpoint, we are actually more interested in the complement of the PVN or 1 – PVN. This measure, which can also be expressed as P(D+|T-), tells the clinician what the probability is of having the disease despite testing negative (i.e., the rate of false negative test results among all negative test results). A high PVN means that there are few false negative results among all test negative results, so an alternative diagnosis should be sought.

Prevalence: Prevalence simply represents the proportion of the total population tested that have disease, or

$$P = \frac{\text{Total Number of Diseased}}{\text{Total Population (N)}} = \frac{TP + FN}{TP+FN+FP+TN} = \frac{a + c}{a + b + c + d}$$

Prevalence is very important since it has a dramatic influence on predictive value positive and negative. Prevalence is the **"third force" -** it is the player that often goes unnoticed only to reveal its influence in dramatic fashion! Other equivalent names for prevalence include the likelihood of disease, prior probability, prior belief, prior odds, pre-test probability and pre-test odds.

Lets go back and look at the example of D-dimer testing and DVT:

Figure 8. The PVP and PVN of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995) Prevalence = 25%

	DVT		
	PRESENT (D+)	**ABSENT (D-)**	
POS (T+)	47 a b c	37 d	PVP = 47/84 = 56%
NEG (T-)	6	124	PVN = 124/130 = 95%
	N = 53	N = 161	N = 214
	Se = 89%	Sp = 77%	

From this 2-by-2 table, we can calculate the PVP as 47/84 = 56% and the PVN as 124/130 = 95%. So, if the test is positive we are only 56% sure that the patient has the disease (about as sure a tossing a coin), whereas if the test is negative we are 95% sure that the subject is disease free. Obviously this test is extremely good for ruling out DVT but practically worthless at ruling in DVT! The other important piece of information to note is the high prevalence of disease in this population i.e., 53/214 = 25%.

The prevalence of DVT in this population was very high, since this group of patients had been admitted to one of 2 Hamilton, Ont., area referral hospitals participating in this research project. Now lets look at the situation that a community-based physician might face. In a typical community-based hospital, the prevalence of DVT in a group of patients complaining of leg pain and swelling is likely to be lower. For illustration, lets say its 5%. The primary care physicians at this community-based hospital are very excited to try out the new test that performed so well in Hamilton. The hospital used the same D-dimer test in another 214 sequential patients with clinical signs consistent with DVT. The Se and Sp of the test are still the same (i.e., 89% and 77%, respectively). The results obtained are shown below:

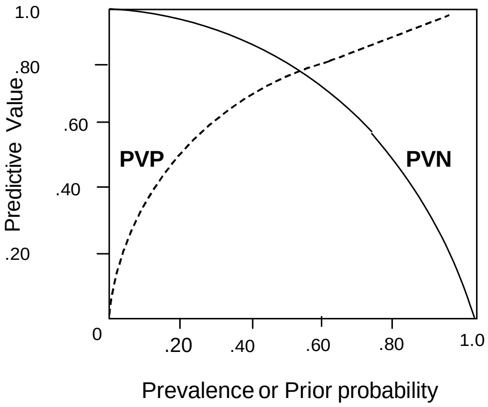
Figure 9. The PVP and PVN of D-dimer whole blood assay (SimpliRED assay) for DVT (Ref: Wells PS, Circulation, 1995). Prevalence of DVT = 5%

	DVT		
	PRESENT (D+)	**ABSENT (D-)**	
POS (T+)	10 a b c	47 d	PVP = 10/57 = 18%
NEG (T-)	1	156	PVN = 156/157 = 99.4%
	N = 11	N = 203	N = 214
	Se = 89%	Sp = 77%	

The physicians are surprised to find out that only 18% of the patients who tested positive actually had DVT. Explanation?.... The lower disease prevalence meant that proportionally more patients were placed in the disease absent column (cells b and d), while fewer patients were placed in the disease present column (cells a and c). So, in the above example, 5% of the 214 subjects (i.e., 11) were put in the disease present column, while 203 were put in the disease absent column. Although the Sp remained at 77%, the
23% FP rate meant that 47 of the 203 patients that did not have DVT had FP results (cell b). The Se also remained the same (89%), so 10 of the 11 patients that had DVT tested positive (cell a). The PVP is lower because the **relative size** of cell a compared to cell b is now much smaller.

One will also note from this example that PVN increased with the lower prevalence - again this is a result of the change in relative sizes of cells c and d. The influence of prevalence on PVP and PVN is demonstrated in the following figure:

Figure 10. The PVP and PVN as a Function of Prevalence for a Typical Diagnostic Test



The effect of prevalence can be summarized as follows:

As prevalence falls, positive predictive value must fall along with it, and negative predictive value must rise. Conversely, as prevalence increases, positive predictive value will increase and negative predictive value will fall.

IV. Bayes’ theorem - the calculation of predictive values for any combination of Se, Sp and Prevalence values using a 2 x 2 table

Bayes theorem is essentially the process by which disease probabilities are revised in face of new test information. We will learn much more about the application of Bayes theorem in diagnostic testing in Epi-547, but for now our task is to be able to calculate predictive values for any combination of Se, Sp and prevalence values using a 2 x 2 table. A step-by-step approach (using an example with Se = 90%, Sp = 80% and prevalence = 10%) follows:

1. First pencil out a 2 x 2 table with disease status (present or absent) along the top (columns) and test status (positive or negative) along the left-hand side (or rows).
2. Next fix the total number of subjects (N) to be included in the table. You can use any number but obviously it makes it easier to use a whole number such as 100 or 1000. Lets pick 1,000.
3. Now calculate the expected number of diseased individuals by applying the disease prevalence rate of 10% to the 1,000 subjects (= 100), and place them at the bottom of the left hand (disease +) column.
4. Place the 900 disease-free subjects at the bottom of the right hand column.
5. Calculate the number of subjects in the top left cell (cell “a”) by multiplying 100 by the sensitivity (i.e., 0.90 x 100 = 90) and place the remaining 10 in the lower left cell (cell “c”) – these are the false negative subjects.
6. Likewise use the Sp of 80% and the 900 subjects to calculate the numbers of subjects in cells “b” and “d” (180 and 720, respectively)
7. Now use the top row (cells a and b) to calculate the PPV (90/270 = 33%)
8. And use the lower row of cells (cells c and d) to calculate the PVN (720/730 = 98.6%). (N.B. you can also calculate PVP and PVN directly using the two Bayes’ equations shown in the lecture).
9. Your table should look like below.
10. Practice doing this using the table of Se, Sp and Prevalence values on D2L.

Disease

	PRESENT (D+)	ABSENT (D-)	
POS (T+)	90	180	PVP = 90/270 = 33%
	a b c	d	
NEG (T-)	10	720	PVN = 720/730 = 98.6%
	N = 100	N = 900	N = 1000
	Se = 90%	Sp = 80%	

V. Multiple testing strategies

Diagnostic tests that have sufficiently high sensitivity and specificity that they can simultaneously “rule out” and “rule-in” are very rare. Generally the physician has access to an array of imperfect tests. However, armed with a good understanding of these diagnostic test principles and Bayes’ theorem, she will be able to squeeze out more information from the available tool box by combining tests. There are two ways of doing this:

Parallel testing

This describes the situation where several test are run simultaneously (i.e., a panel of tests) and any one positive test leads onto further evaluation (see Figure 3.12 in FF). The net effect is to increase the likelihood of detecting disease, so sensitivity increases (because there are multiple opportunities for a positive test result), as does PVN. However, as is always the case, there is a price to pay - the probability of false positive results increases - so both specificity and PVP decline. Parallel testing is typically used in the early phases of the work up where you are trying to quickly rule out several

conditions – by running a panel of related tests - if they all come back negative - then the condition(s) can be “ruled out” (think SnOut – parallel testing maximizes Se and so PVN is maximized). However, when using this strategy a positive test means little (other than more testing is required) because the increase false positive results in lower PVP. The parallel testing strategy can be easily abused by using it as a screening tool for “anything and everything”. This approach, often favoured by neophyte interns and residents, is very costly, highly inefficient, dangerous to the patient (who has to undergo unnecessary follow-up tests because of the false positive results), and is ultimately bad medicine. As explained in the FF text (page 53-55) this strategy works best when a highly sensitive test strategy is required but you are armed with 2 or more relative insensitive tests. If these tests measure different clinical phenomenon (i.e., they provide independent information), then combining them in parallel maximizes your chance of identifying diseased subjects.

Serial testing

This describes the situation where several tests are run in order or series and each subsequent test is only run if the first test was positive (see Figure 3.12 in FF). In this approach any negative test leads to the suspension of the work-up. The net effect is to increase specificity and PVP because each case has to test positive to multiple tests (so false positives are rare). However, again there is a price to pay - the probability of false negative results increases so sensitivity declines as dose PVN. Serial testing is typically used when one wants to be sure that a disease is ruled in with certainty, and there is no rush to do so. It is also used when a particular definitive test is expensive, difficult, or invasive – to avoid over-using such a test, a cheaper and/or less invasive test is run first and only those testing positive go on to have the definitive test. An example would be the use of a colonoscopy following a positive fecal occult blood test. Finally, the use of serial independent tests is a great example of the logic of Bayes’ theorem to revise

probabilities. The results of the first test are used to provide the pre-test probability for the second test - see Fig 3.13 in FF which shows an example using likelihood ratios (a topic for Epi-547).

EPI-546 Block I

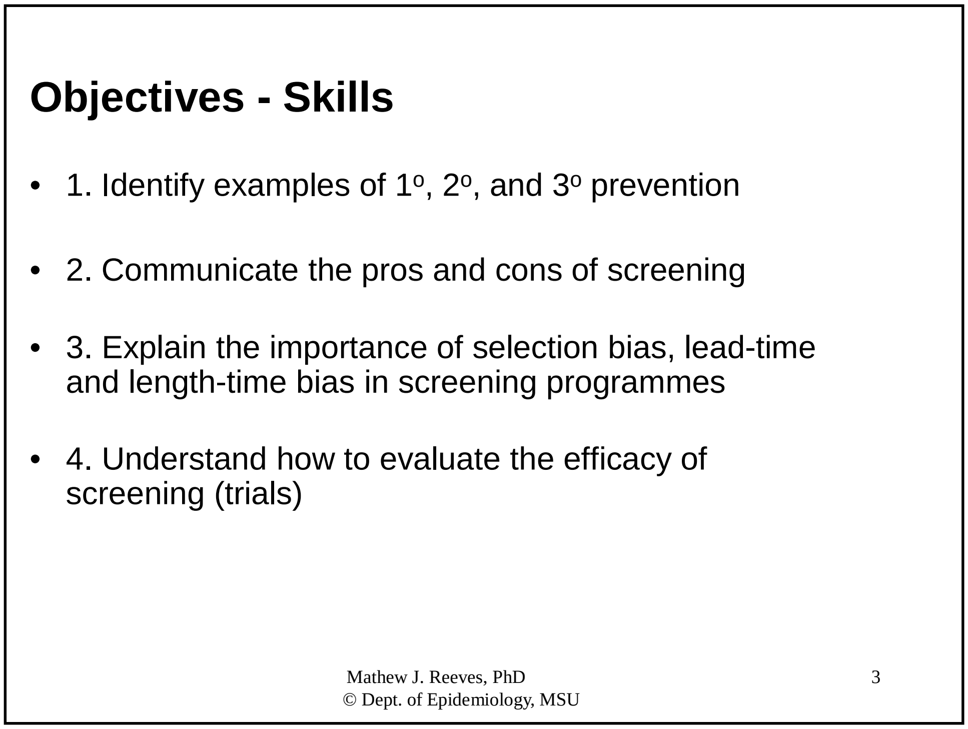
Lecture – Prevention

Understanding the rationale for disease prevention and early intervention (screening)

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology

Objectives - Concepts

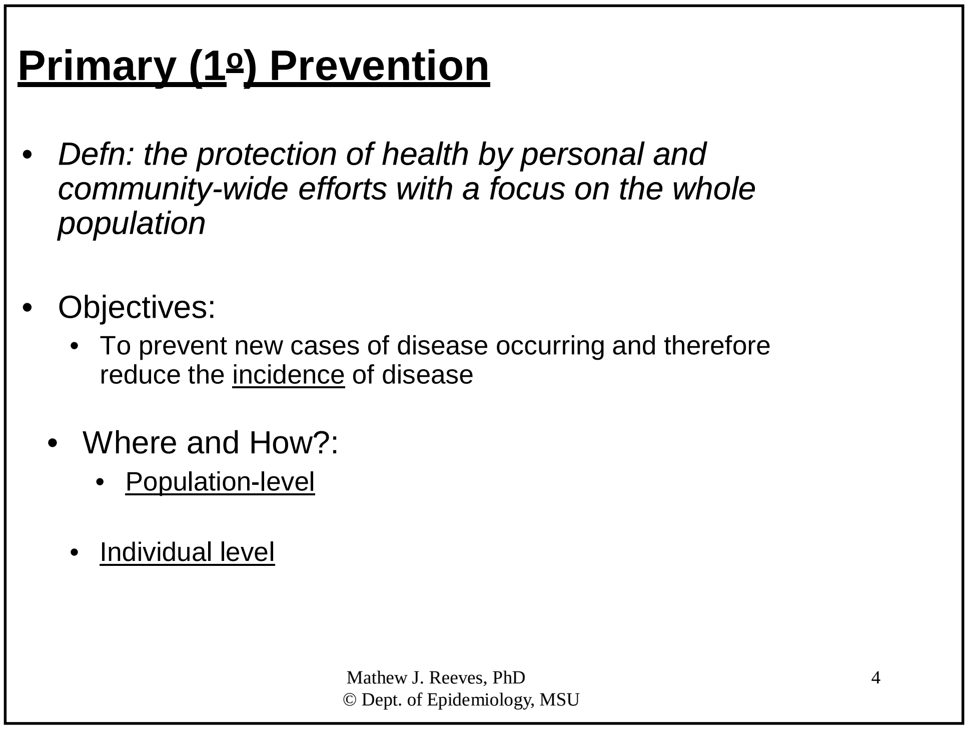
- 1. Primary (1^o), Secondary (2^o), and Tertiary (3^o) prevention
- 2. Population-level vs. individual-level prevention
- 3. Screening (secondary prevention)
  - Mass screening vs. case-finding
- 4. Screening concepts
  - Pre-clinical phase, lead time, test Se & Sp, importance of trials, DSMR
- 5. Screening Biases (Observational studies)
  - Lead-time, Length-time, and Compliance
- 6. Assessing the feasibility of screening
- 7. Risks (Harms) vs. Benefits



Objectives - Skills

- 1. Identify examples of 1^{o}, 2^{o}, and 3^{o} prevention
- 2. Communicate the pros and cons of screening
- 3. Explain the importance of selection bias, lead-time and length-time bias in screening programmes
- 4. Understand how to evaluate the efficacy of screening (trials)

Mathew J. Reeves, PhD 3
© Dept. of Epidemiology, MSU



Primary (1^{o}) Prevention

- *Defn: the protection of health by personal and community-wide efforts with a focus on the whole population*
- Objectives:
  - To prevent new cases of disease occurring and therefore reduce the incidence of disease
- Where and How?:
  - Population-level
  - Individual level

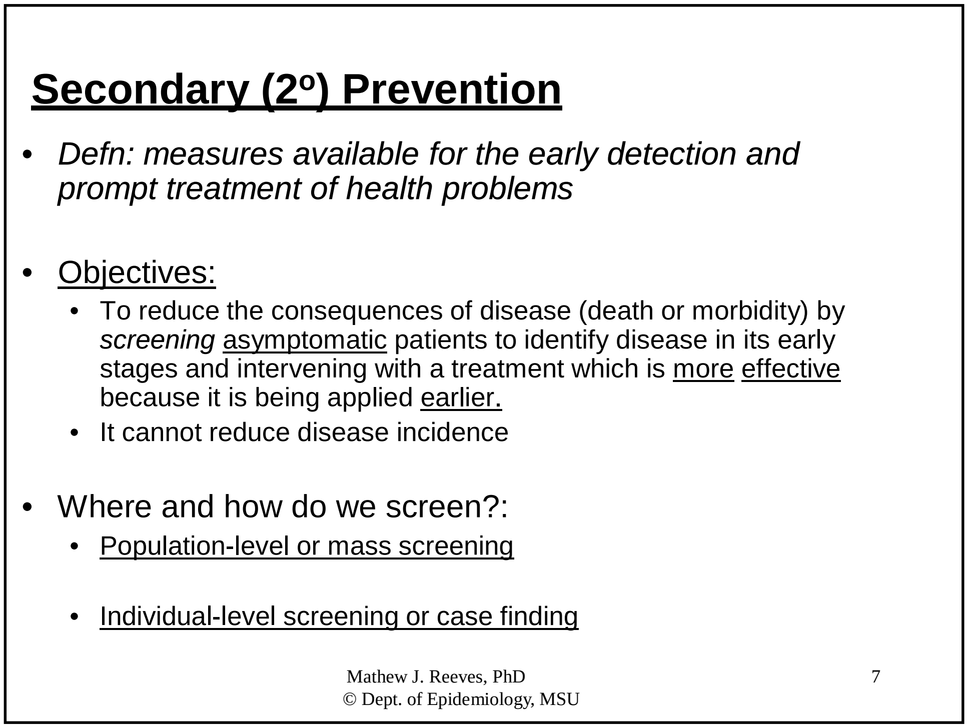
Mathew J. Reeves, PhD 4
© Dept. of Epidemiology, MSU

1^{o} Prevention @ Population-level

- By reducing exposure to causal (risk) factors
  - e.g., reducing smoking initiation in teenagers
- By adding a factor that prevents disease
  - e.g., vaccination, water fluoridation
- Usually requires policy and/or legislation
  - Smoking = tobacco taxes, restrictions on smoking indoors
  - Physical activity = structural changes to the environment
    - sidewalks, walking paths, bike lanes (Town planning)
- Primary prevention at the population-level works best when it is driven by changes in societal attitudes
  - e.g., drinking and driving, bikes lanes

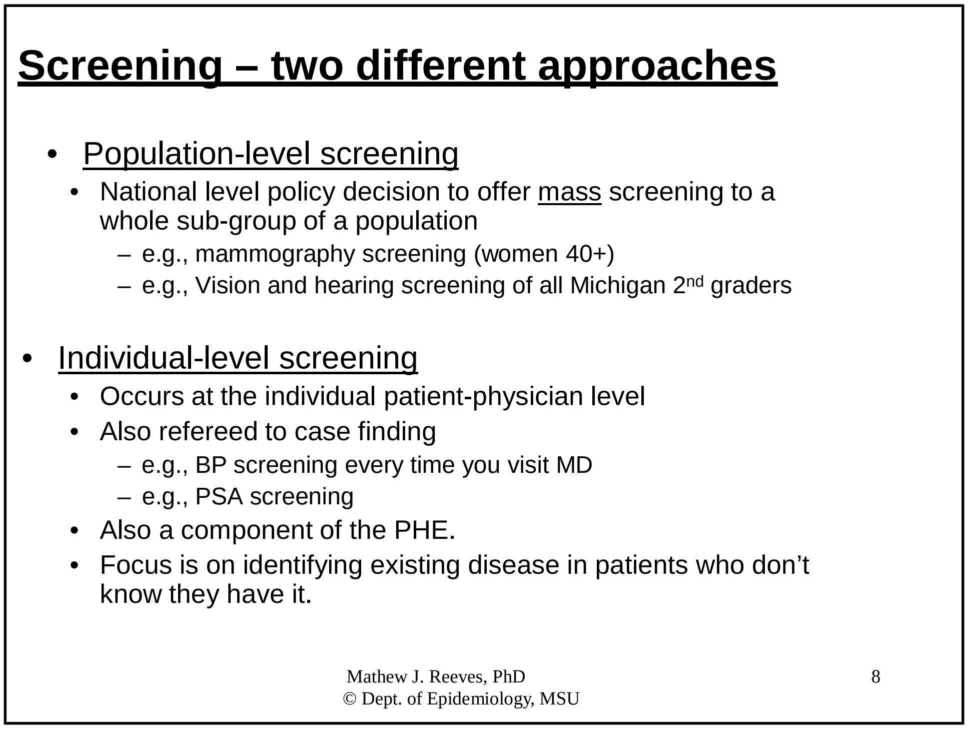
1^{o} Prevention @ Individual-level

- By removing or lowering risk factors in at-risk patients
  - Occurs at the patient-physician level
  - Rationale behind the periodic health exam (PHE)
    - e.g., smoking cessation counseling
    - e.g., risk factor screening in at-risk patients (BP, BC, Physical inactivity, abdominal obesity)
- Distinction between primary prevention and secondary prevention hinges on the presence of existing disease
  - Primary prevention = no existing disease



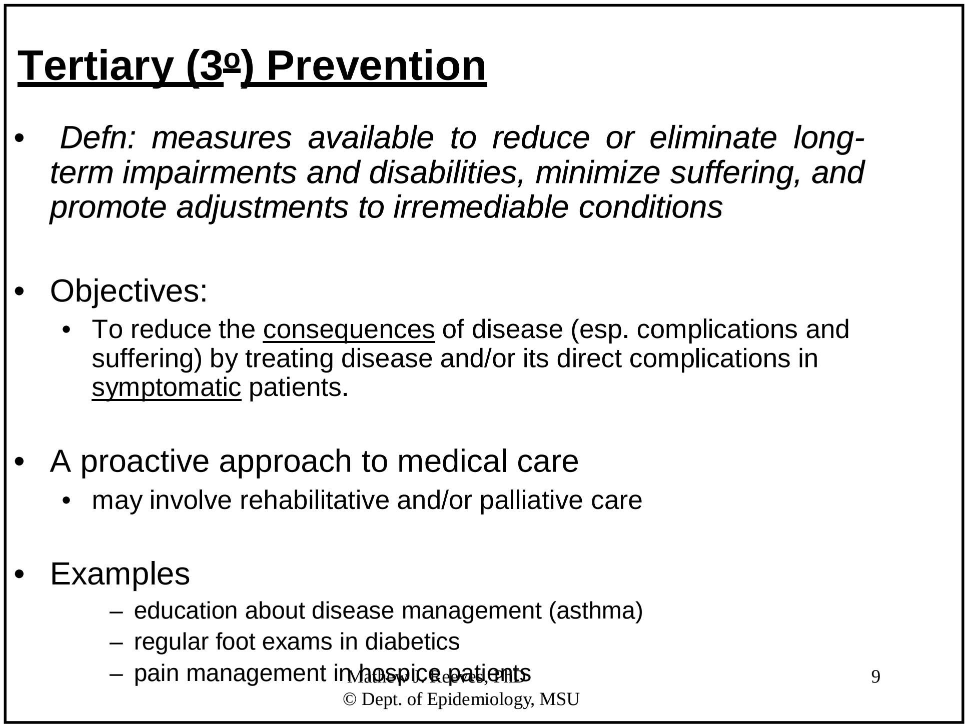
Secondary (2^o) Prevention

- *Defn: measures available for the early detection and prompt treatment of health problems*
- Objectives:
  - To reduce the consequences of disease (death or morbidity) by *screening* asymptomatic patients to identify disease in its early stages and intervening with a treatment which is more effective because it is being applied earlier.
  - It cannot reduce disease incidence
- Where and how do we screen?:
  - Population-level or mass screening
  - Individual-level screening or case finding



Screening – two different approaches

- Population-level screening
  - National level policy decision to offer mass screening to a whole sub-group of a population
    - e.g., mammography screening (women 40+)
    - e.g., Vision and hearing screening of all Michigan 2nd graders
- Individual-level screening
  - Occurs at the individual patient-physician level
  - Also refereed to case finding
    - e.g., BP screening every time you visit MD
    - e.g., PSA screening
  - Also a component of the PHE.
  - Focus is on identifying existing disease in patients who don’t know they have it.



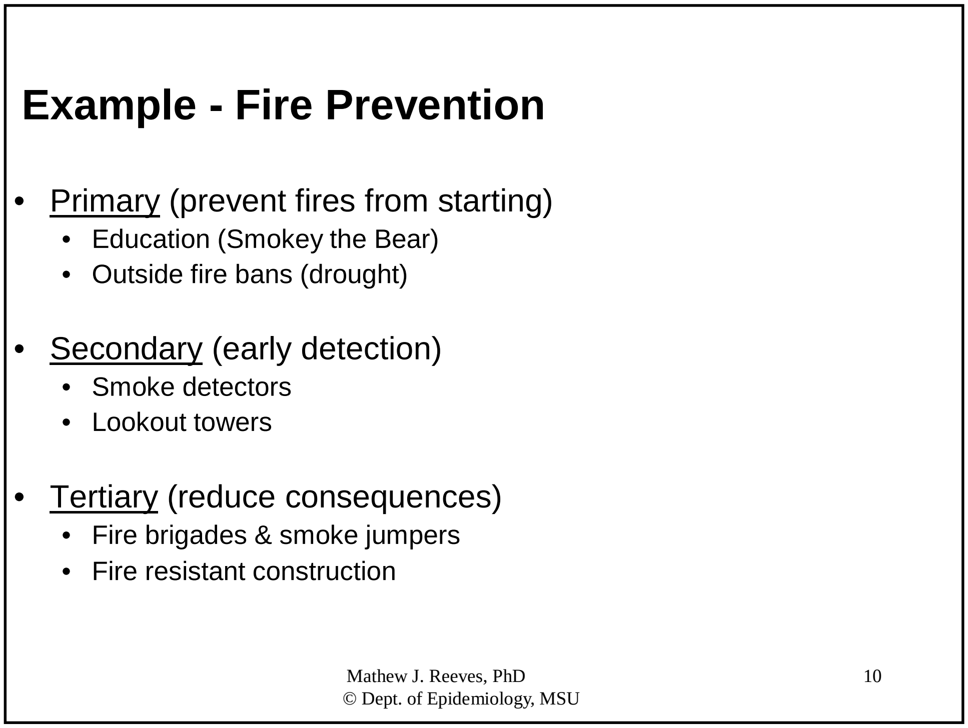
Tertiary (3^o) Prevention

- *Defn: measures available to reduce or eliminate long-term impairments and disabilities, minimize suffering, and promote adjustments to irremediable conditions*

- Objectives:
  - To reduce the consequences of disease (esp. complications and suffering) by treating disease and/or its direct complications in symptomatic patients.

- A proactive approach to medical care
  - may involve rehabilitative and/or palliative care

- Examples
  - – education about disease management (asthma)
  - – regular foot exams in diabetics
  - – pain management in hospice patients

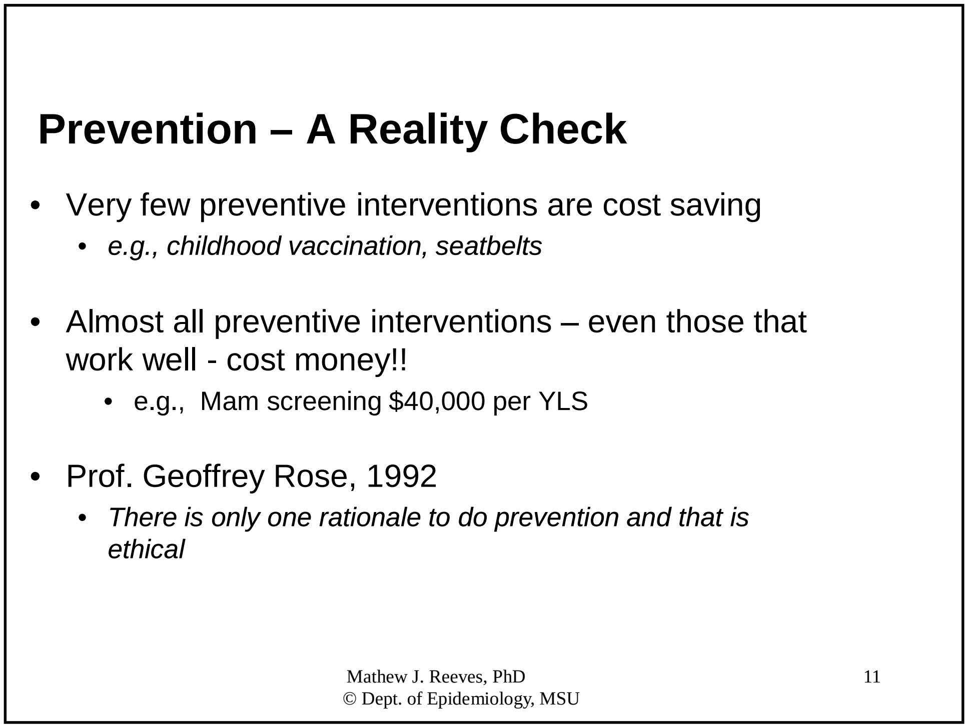


Example - Fire Prevention

- Primary (prevent fires from starting)
  - Education (Smokey the Bear)
  - Outside fire bans (drought)

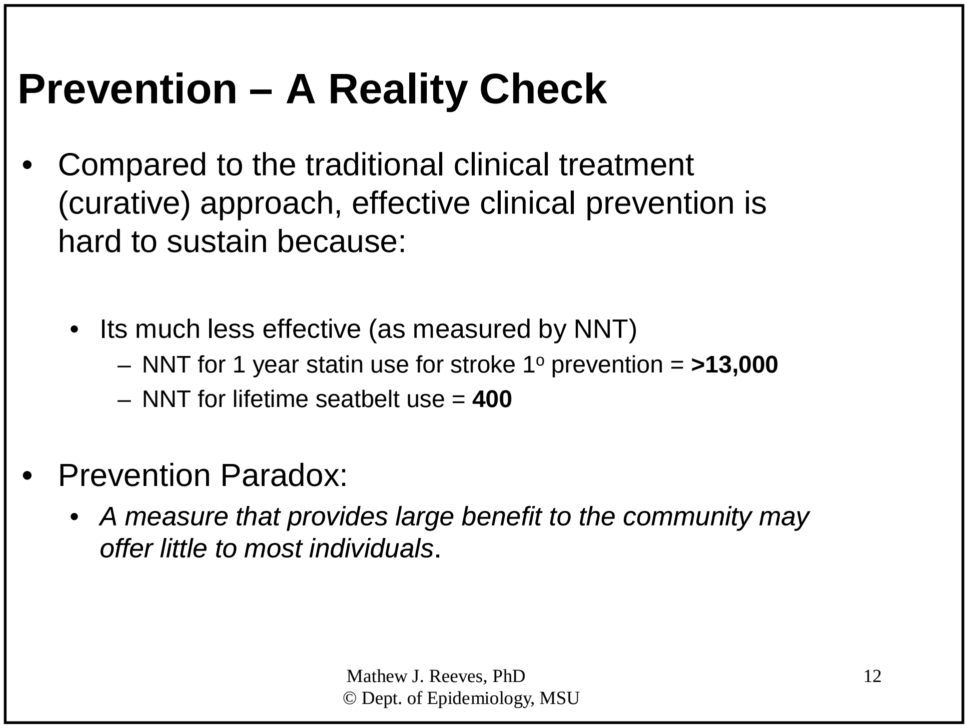
- Secondary (early detection)
  - Smoke detectors
  - Lookout towers

- Tertiary (reduce consequences)
  - Fire brigades & smoke jumpers
  - Fire resistant construction



Prevention – A Reality Check

- Very few preventive interventions are cost saving
  - *e.g., childhood vaccination, seatbelts*
- Almost all preventive interventions – even those that work well - cost money!!
  - e.g., Mam screening $40,000 per YLS
- Prof. Geoffrey Rose, 1992
  - *There is only one rationale to do prevention and that is ethical*

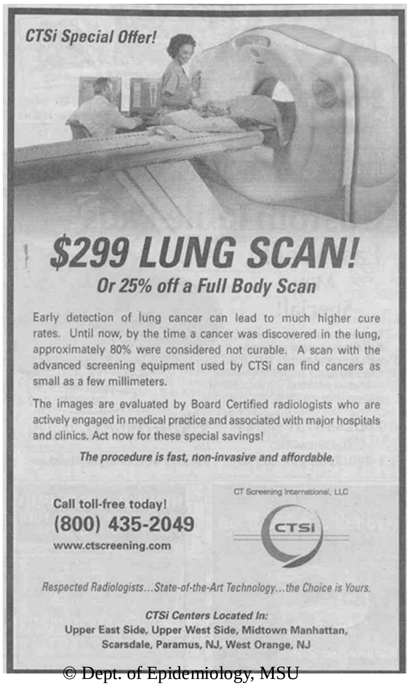


Prevention – A Reality Check

- Compared to the traditional clinical treatment (curative) approach, effective clinical prevention is hard to sustain because:
  - Its much less effective (as measured by NNT)
    - NNT for 1 year statin use for stroke 1^{o} prevention = **>13,000**
    - NNT for lifetime seatbelt use = **400**
- Prevention Paradox:
  - *A measure that provides large benefit to the community may offer little to most individuals*.

Screening - Introduction

- Objective: to reduce mortality and/or morbidity by early detection and treatment.
- *Secondary prevention*.
- Asymptomatic individuals are classified as either unlikely or possibly having disease.
- Important distinction between mass or population-based screening and case finding.
- The allure of screening brought on by new technology is almost irresistible…..



CTSi Special Offer!

$299 LUNG SCAN!

Or 25% off a Full Body Scan

Early detection of lung cancer can lead to much higher cure rates. Until now, by the time a cancer was discovered in the lung, approximately 80% were considered not curable. A scan with the advanced screening equipment used by CTSi can find cancers as small as a few millimeters.

The images are evaluated by Board Certified radiologists who are actively engaged in medical practice and associated with major hospitals and clinics. Act now for these special savings!

The procedure is fast, non-invasive and affordable.

Call toll-free today!
(800) 435-2049
www.ctscreening.com

CT Screening International, LLC
CTSi

Respected Radiologists…State-of-the-Art Technology…the Choice is Yours.

CTSi Centers Located In:
Upper East Side, Upper West Side, Midtown Manhattan,
Scarsdale, Paramus, NJ, West Orange, NJ

Screening - Introduction

- Effective screening involves both *diagnostic* and *treatment* components

- Screening differs from diagnostic testing:

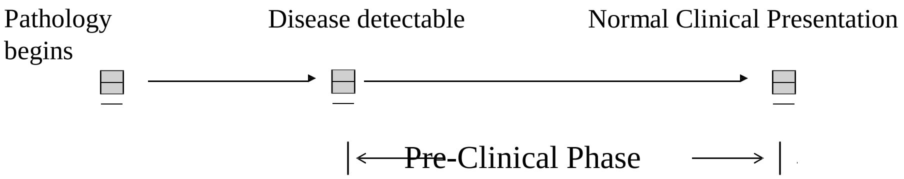
Screening	Testing
Healthy non-patients	Sick patients
No diagnostic intent	Diagnostic intent
Very low to low disease prevalence	Low to high disease prevalence

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

I. Important Concepts in Screening

The Pre-Clinical Phase (PCP)

- the period between when early detection by screening is possible and when the clinical diagnosis would normally have occurred.



Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

Pre-clinical Phase (PCP)

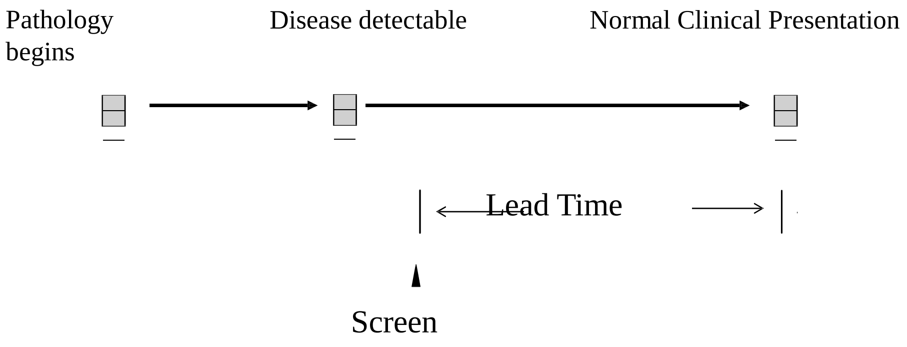
- Important to know PCP since it helps determine:
  - Expected utility of screening
    - Colorectal cancer = 7-10 years
    - Childhood diabetes = 2-6 months
  - Required minimal frequency of screening
    - Mam screening women 40-49 = 1-2 years
    - Mam screening women 50-69 = 3-4 years

- Prevalence of PCP indicates how much early disease there is to detect

- Prevalence of PCP is affected by:
  - disease incidence, average duration of the PCP, previous screening, sensitivity of the test
  - …..see concept of Prevalence pool (Lecture 3)

Lead Time

Lead time = amount of time by which diagnosis is advanced or made earlier



Pathology begins

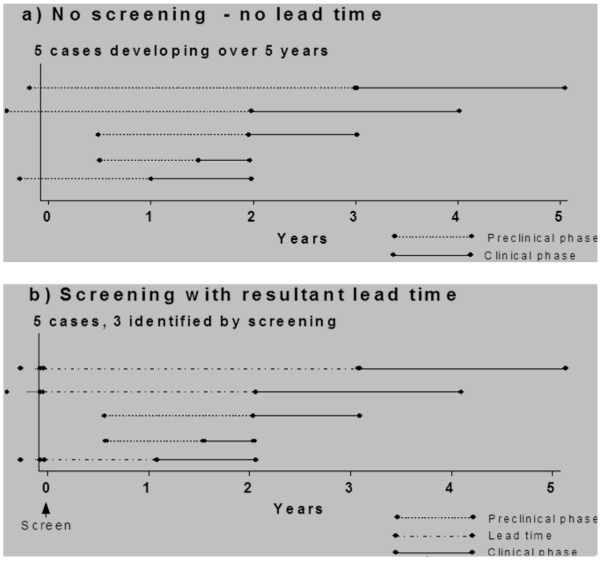
Disease detectable

Normal Clinical Presentation

Lead Time

Screen

Fig 1. Relationship between screening, pre-clinical phase, clinical phase and lead time


a) No screening - no lead time

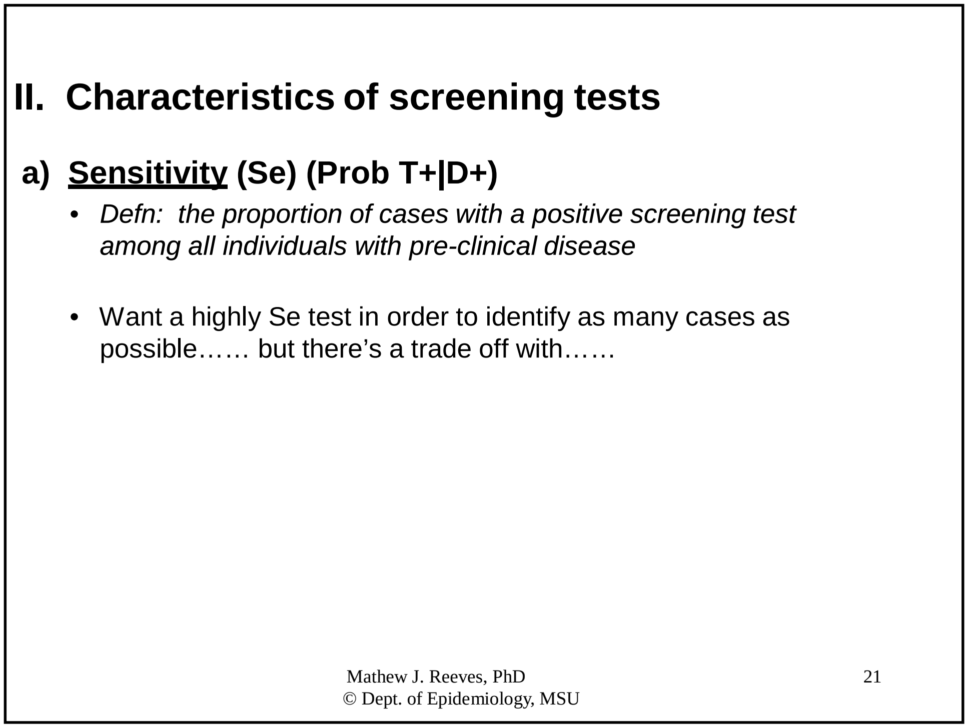
b) Screening with resultant lead time

(Survival = 2yrs)

('Survival' = 5yrs)

Lead Time

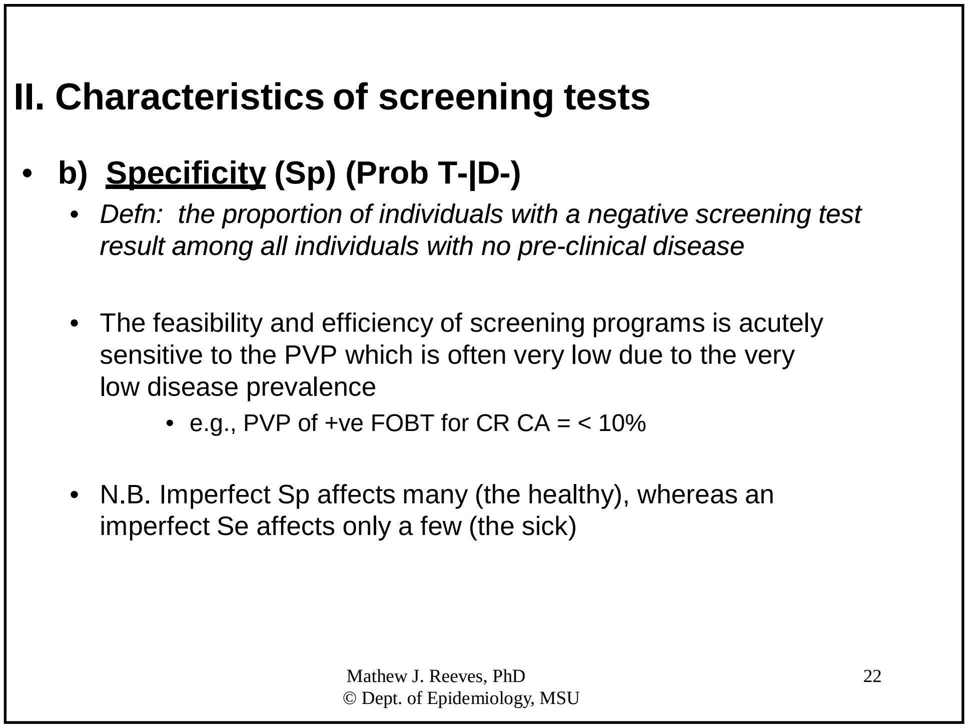
- Equals the amount of time by which treatment is advanced or made “early”
- Not a theory or statistical artifact but what is expected and must occur with early detection
- Does not imply improved outcome!!
- *Necessary but not sufficient condition* for effective screening.



II. Characteristics of screening tests

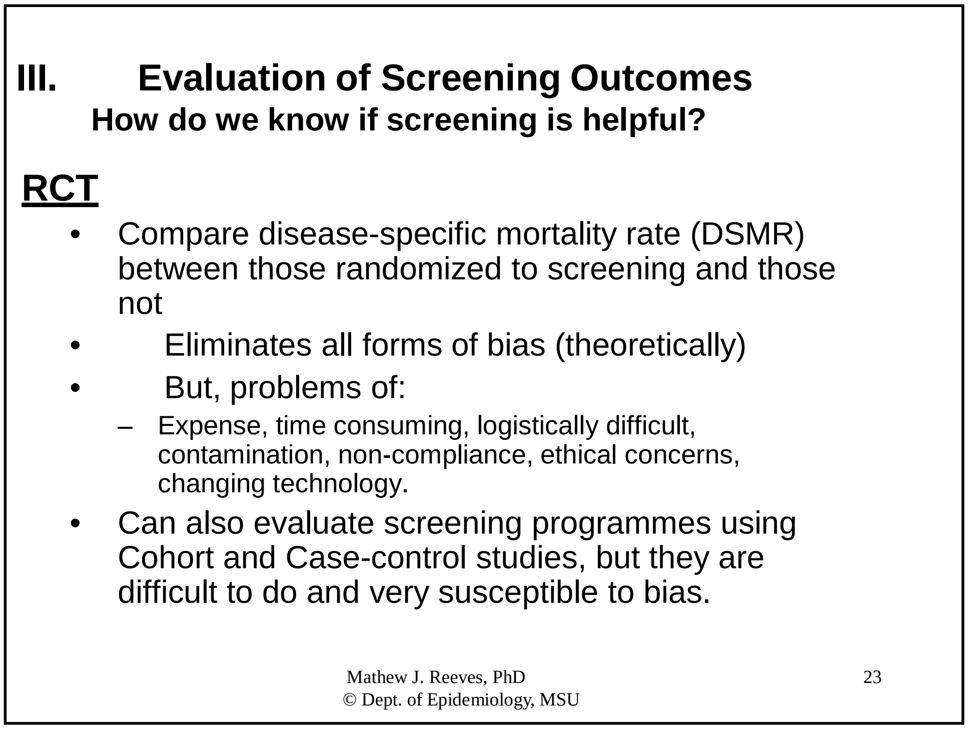
a) Sensitivity (Se) (Prob T+|D+)

- *Defn: the proportion of cases with a positive screening test among all individuals with pre-clinical disease*
- Want a highly Se test in order to identify as many cases as possible…… but there's a trade off with……



II. Characteristics of screening tests

- **b) Specificity (Sp) (Prob T-|D-)**
  - *Defn: the proportion of individuals with a negative screening test result among all individuals with no pre-clinical disease*
  - The feasibility and efficiency of screening programs is acutely sensitive to the PVP which is often very low due to the very low disease prevalence
    - e.g., PVP of +ve FOBT for CR CA = < 10%
  - N.B. Imperfect Sp affects many (the healthy), whereas an imperfect Se affects only a few (the sick)

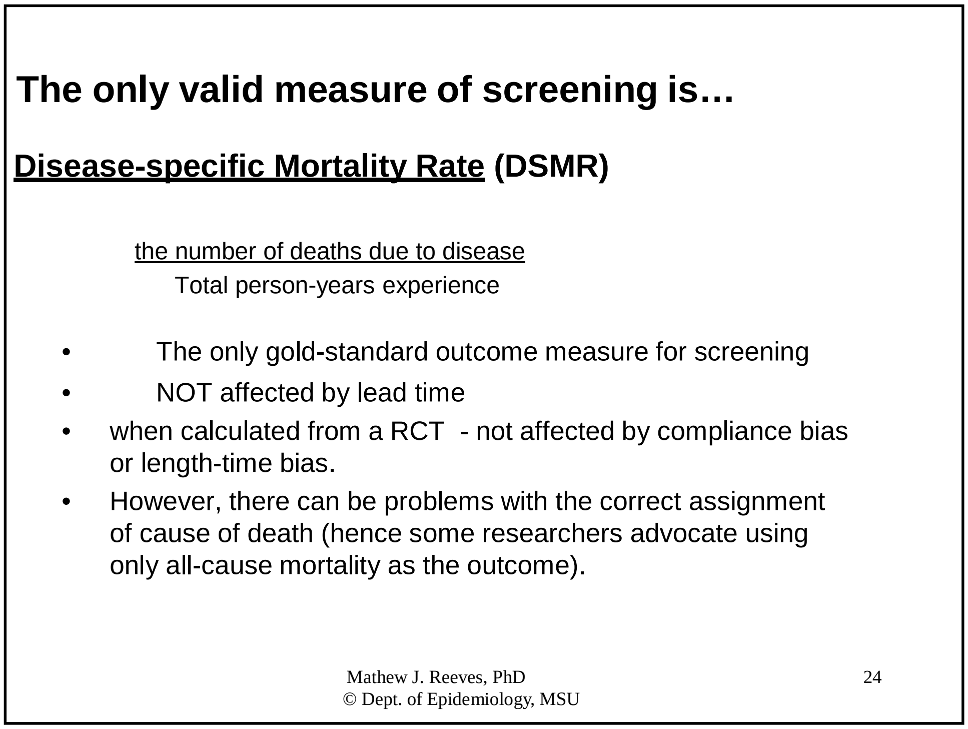


III. Evaluation of Screening Outcomes

How do we know if screening is helpful?

<u>RCT</u>

- Compare disease-specific mortality rate (DSMR) between those randomized to screening and those not
- Eliminates all forms of bias (theoretically)
- But, problems of:
  - Expense, time consuming, logistically difficult, contamination, non-compliance, ethical concerns, changing technology.
- Can also evaluate screening programmes using Cohort and Case-control studies, but they are difficult to do and very susceptible to bias.



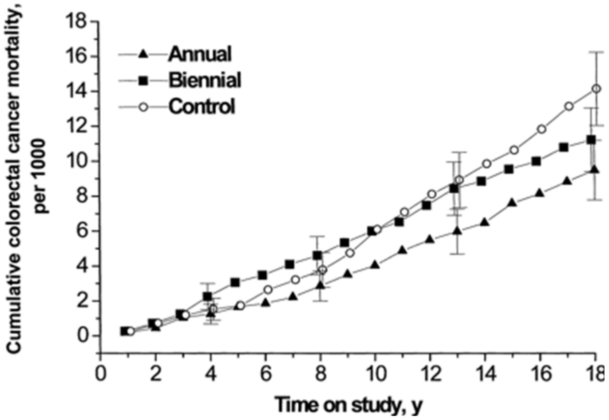
The only valid measure of screening is…

<u>Disease-specific Mortality Rate</u> (DSMR)

$$\frac{\text{the number of deaths due to disease}}{\text{Total person-years experience}}$$

- The only gold-standard outcome measure for screening
- NOT affected by lead time
- when calculated from a RCT - not affected by compliance bias or length-time bias.
- However, there can be problems with the correct assignment of cause of death (hence some researchers advocate using only all-cause mortality as the outcome).

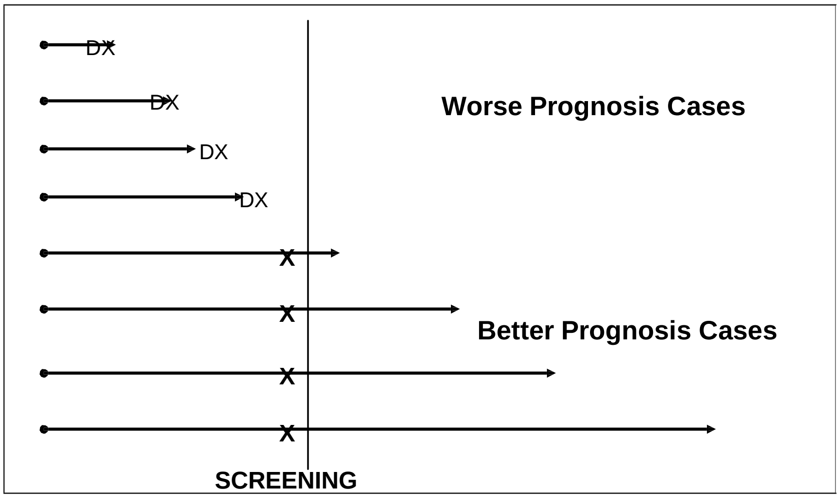
Example of a RCT reporting DSMR to measure efficacy of FOBT screening on Colorectal CA Mortality (Mandel, NEJM 1999)



IV. Biases that effect screening studies

- Observational studies and especially survival data are acutely sensitive to:
- 1. <u>Compliance bias</u> (Selection bias):
  - Volunteers or compliers are better educated and more health conscious – thus they have inherently better prognosis
- 2. <u>Lead-time bias</u>
  - <u>Apparent</u> increased survival duration introduced by the <u>lead</u> <u>time</u> that results from screening.
  - Screen-detected cases survive longer event without benefit of early treatment (review Fig 2 in course notes).
- 3. <u>Length-time bias</u>
  - Screening preferentially identifies slower growing or less progressive cases that have a better prognosis.

Length-time bias – cases with better prognosis detected by screening



Worse Prognosis Cases

Better Prognosis Cases

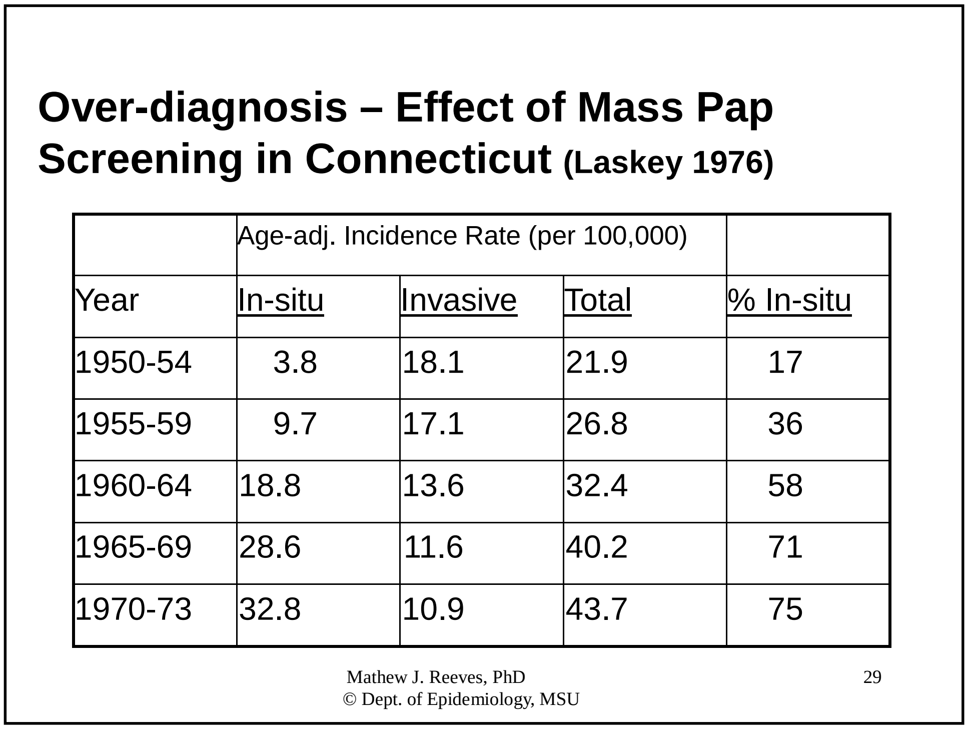
SCREENING

V. Pseudo-disease and Over-diagnosis

- <u>Over-diagnosis</u>
  - Limited malignant potential
  - Extreme form of length-biased sampling
  - Examp: Pap screening and cervical carcinoma

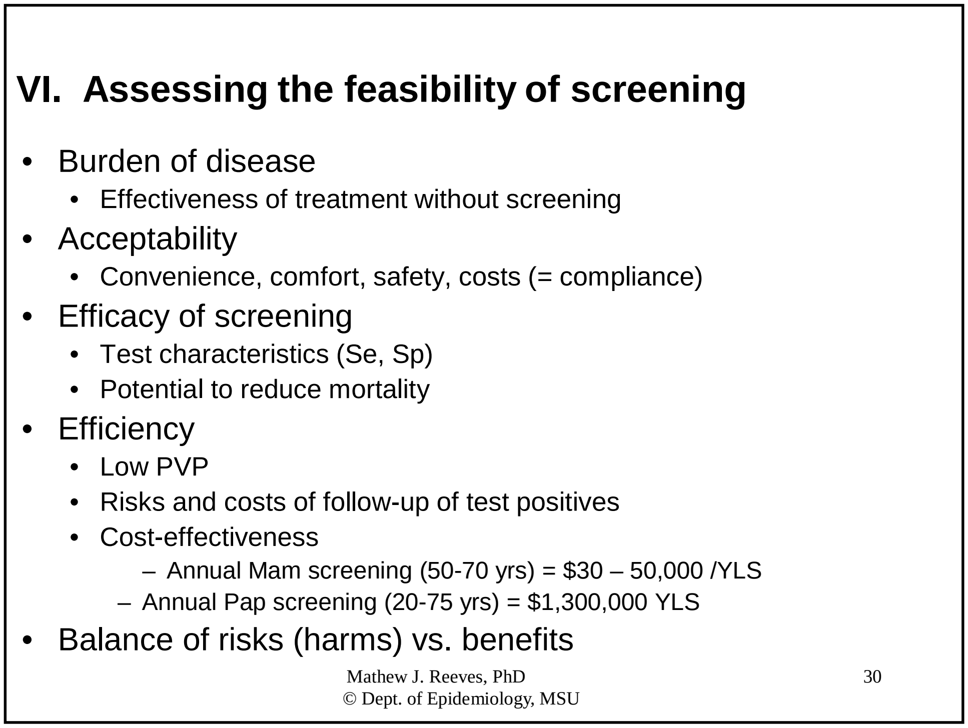
- <u>Competing risks</u>
  - Cases detected that would have been interrupted by an unrelated death
  - Examp: Prostate CA and CVD death

- <u>Serendipity</u>
  - Chance detection due to diagnostic testing for another reason
  - Examp: PSA and prostate CA, FOBT and CR CA



Over-diagnosis – Effect of Mass Pap Screening in Connecticut (Laskey 1976)

	Age-adj. Incidence Rate (per 100,000)			
Year	In-situ	Invasive	Total	% In-situ
1950-54	3.8	18.1	21.9	17
1955-59	9.7	17.1	26.8	36
1960-64	18.8	13.6	32.4	58
1965-69	28.6	11.6	40.2	71
1970-73	32.8	10.9	43.7	75



VI. Assessing the feasibility of screening

- Burden of disease
  - Effectiveness of treatment without screening
- Acceptability
  - Convenience, comfort, safety, costs (= compliance)
- Efficacy of screening
  - Test characteristics (Se, Sp)
  - Potential to reduce mortality
- Efficiency
  - Low PVP
  - Risks and costs of follow-up of test positives
  - Cost-effectiveness
    - Annual Mam screening (50-70 yrs) = $30 – 50,000 /YLS
    - Annual Pap screening (20-75 yrs) = $1,300,000 YLS
- Balance of risks (harms) vs. benefits

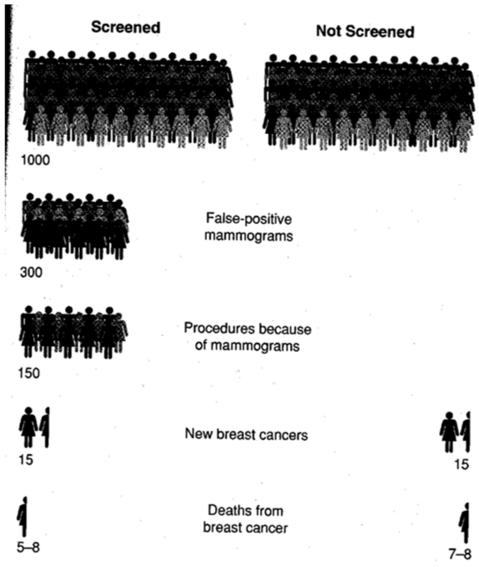


Figure 8.8. Weighing benefit and harm from screening. What happens during a decade of annual mammography in 1000 women starting at age 40.

Feasibility

- Three questions to ask before screening:

• <u>Efficacy</u>	• Should we screen? (scientific)
• <u>Effectiveness</u>	• Can we screen? (practical)
• <u>Cost-effectiveness</u>	• Is it worth it? (scientific, practical, policy, political)

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes – Prevention

Mat Reeves BVSc, PhD

Objectives

1. Identify and distinguish between Primary (1^o), Secondary (2^o), and Tertiary (3^o) prevention
2. Understand the difference between population-level vs. individual-level prevention and identify different examples
3. Understand the role of screening as a form of secondary prevention, and distinguish between mass screening and case-finding
4. Define, understand and apply the following key screening concepts:
   - Pre-clinical phase, lead time, test Se & Sp
5. Understand the importance of randomized trials and the role of the DSMR in determining the value of screening
6. Understand and identify the biases that occur in observational studies of screening:
   - Lead-time, Length-time, and Compliance
7. Understand and be able to apply the criteria used to assess the feasibility of screening
8. Understand the importance of balancing the harms versus benefits of screening at the individual and population level.

Outline:

I. Introduction

The goals of screening are to reduce mortality and morbidity (and/or avoiding expensive or toxic treatments). Screening is a form of *secondary prevention.* Screening is designed to detect disease early in its asymptomatic phase whereby early treatment then either slows the progression of disease or provides a cure. The premise of screening is based on concept that early treatment will stop or retard progression of disease. Screening therefore has both diagnostic and therapeutic components.

Screening involves the examination of asymptomatic people who are then classified as either:

- Unlikely to have disease (TN or FN), or

- Likely to have disease and therefore require further diagnostic evaluation.

Screening is very different from diagnostic testing:

Testing	Screening
Sick patients are tested	Healthy, non-patients are screened
Diagnostic intent	No diagnostic intent
Low to high disease prevalence	Very low to low disease prevalence

There are two fundamentally different types of screening:

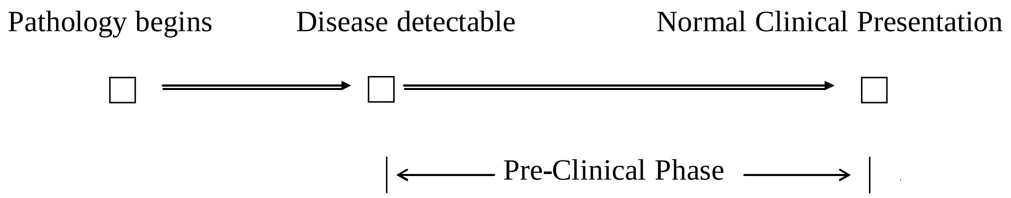
Mass or population-based screening is the application of screening tests to large, unselected populations e.g., mammography screening for breast cancer in women < 40 yrs of age.

Case finding is the use of screening by clinicians to identify disease in patients who present for other unrelated problems e.g., blood pressure measurements.

The format, organization, and intent of these two types of screening are fundamentally different. Mass screening requires a completely different organizational approach to successfully implement - involving policy makers, government, the medical community, and public health on a national basis.

II. Characteristics of Disease

For a disease to be a suitable candidate for screening it must have a sufficiently long *pre-clinical phase*. Pre-clinical phase is defined as the period between when early detection by screening is possible and when the clinical diagnosis would usually have been made.



A. **Pre-clinical phase (PCP):**

The point that a typical person seeks medical attention depends upon availability of medical care, as well as the level of medical awareness in the population. An example of disease with a long pre-clinical phase that suggests screening might be useful is colorectal cancer (i.e., PCP = 7-10 years). Diseases with a short pre-clinical phase are unlikely to be good candidates for screening e.g., childhood diabetes (weeks to a few months).

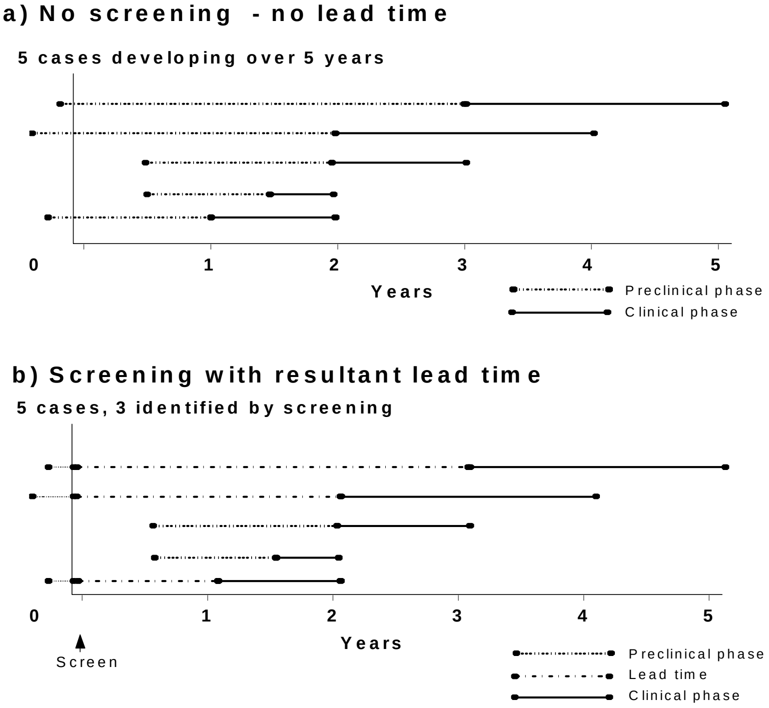
The prevalence of detectable pre-clinical disease in a population (and NOT the prevalence of disease itself) is a critical determinant of the potential utility of screening. The prevalence of pre-clinical disease is dependent upon:

i incidence rate of disease
ii. average length (duration) of pre-clinical phase
iii. recent screening (decreases prevalence)
iv. detection capabilities of the test (greater sensitivity results in higher prevalence)

III. Lead time

Lead time is the interval from detection by screening to the time at which diagnosis would have been made without screening. Lead time is the central rational of screening since it equals the amount of time by which treatment is advanced or made "early". Lead time results in the longer awareness of the disease and does not necessarily imply any improved outcome (since after lead time has occurred early treatment must then be effective for screening to be beneficial). Figure 1 illustrates this concept, in panel a (no screening), there are three cases at time zero who are already in the PCP i.e., in whom pathology has already started. In panel b, screening is conducted at time 0 and 'converts' the PCP in these three subjects into lead time. Thus the disease is advanced 3 years, 2 years, and 1 year, respectively, in these 3 cases.

Figure 1. Relationship Between Screening, Preclinical Phase, Clinical Phase and Lead Time



Lead time is not a theory or statistical artifact, it is what would be expected with early diagnosis and what *must* occur if screening is to be worthwhile (it is therefore a *necessary but not sufficient* condition for screening to be effective in reducing mortality).

Knowledge of the distribution of lead times is useful because it indicates the length of time by which detection and treatment must be advanced in order to achieve a level of improved mortality. It can also help suggest how often you should screen - for example, the estimated lead time for invasive colo-rectal cancer is 7-10 years. Thus guidelines suggest that an appropriate screening interval for sigmoidoscopy is every 5 years.

IV. Characteristics of screening tests

A. ***Sensitivity***: the proportion of cases with a positive screening test among all cases of pre-clinical disease.

The target disorder is the pre-clinical lesion, not clinically evident disease. Test operating characteristics maybe very different between the two. Se is often first determined by applying tests to symptomatic patients, but screening Se is likely to be lower. For sensitivity to be accurately defined, all individuals who have the pre-clinical disease must be identified using an acceptable "gold standard" diagnostic test. However, the true disease status of individuals who have a negative screening test is impossible to verify, since there is no justification to do a full diagnostic work-up (this is an excellent example of *verification bias*). Usually, Se can only be estimated in screening studies by counting the number of *interval cases* that occur over a specified period (e.g., 12 months) in persons who tested negative to the screening test. These interval cases are regarded as false negatives (FNs).

B. ***Specificity***: ability of screening test to designate as negative people who do not have pre-clinical disease.

There is always an inherent trade-off between Se and Sp - as one increases the other must decline. Note also that an imperfect Se affects a few (the cases), whereas an imperfect Sp affects many (the healthy!). The FP rate (1 - Sp) needs to be sufficiently low for screening to be feasible, because the prevalence of pre-clinical disease is always low, thus the *predictive value positive (PVP)* will be low in most screening programs. (PVP is defined as the proportion of screen positive subjects who have pre- clinical disease i.e., TP/(TP + FP)).

PVP can be improved by screening only high risk populations or using a lower frequency of screening (which increases prevalence of pre-clinical disease). Because the prevalence of pre-clinical disease will fall in populations that are repeatedly screened, PVP will be expected to decline in a successful screening program making it increasingly inefficient.

C. ***Yield*****:** the amount of previously unrecognized disease that is diagnosed and brought to treatment as a result of screening.

Yield is affected by the Se of the screening test (a lower Se means that a smaller fraction of diseased individuals are detected at any screening), and the prevalence of pre-clinical disease in the population. A higher prevalence will increase the yield, thus aiming screening programs at high risk populations will increase its efficiency.

V. Evaluation of screening outcomes

A. ***Methods**:*

Experimental: Conduct a RCT of the screening modality and compare the disease- specific cumulative mortality rate between the groups randomized to screening or usual care (control). The randomized design is critical in eliminating confounding due to unknown and known factors, and in allowing a valid comparison (unaffected by lead time bias) between the two groups. The RCT also allows one to study the effects of early treatment, to estimate the distribution of lead times, and identify prognostic factors.

Problems: Expense, time (many years before results are available by which time the screening technology has often changed logistical problems, ethical concerns.

Non-experimental:

I. Cohort - comparison of advanced illness or death rate between people who chose to be screened and those that do not.
II. CCS - comparison of screening history between people with advanced disease (or death) and those unaffected (healthy).
III. Ecological - correlation of screening patterns and disease experience of several populations.

Problems:

I. Confounding due to "health awareness" (people who choose to get screened are more health conscious and have lower mortality).
II. Poor quality, often retrospective data.
III. Difficult to distinguish screening from diagnostic examinations.

B. ***Measures of effect:***

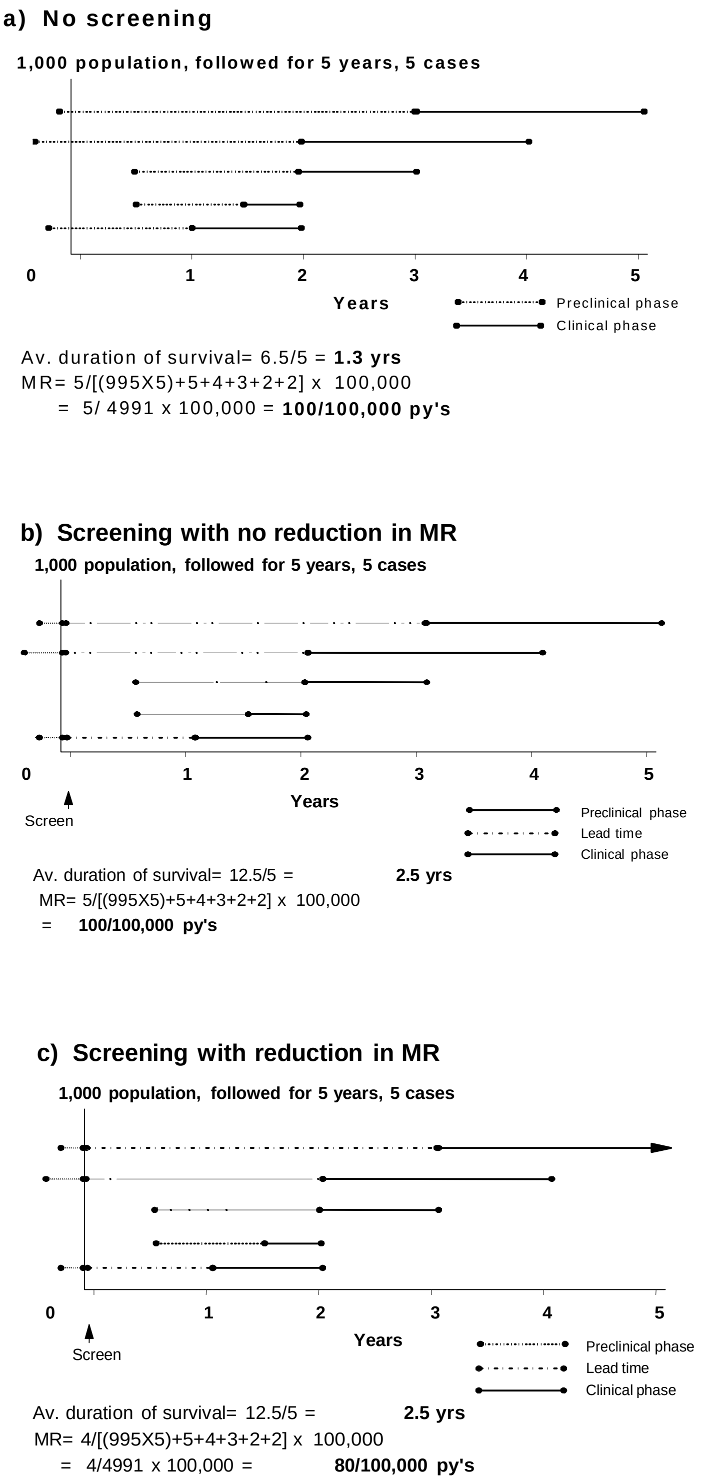
1) Comparison of survival experience (or duration)

Important!!! The efficacy of a screening program cannot be assessed by comparing the duration of survival of screen detected cases and cases diagnosed clinically. Despite the fact that this is commonly done, such analyses over-estimate the effect of screening because of the following three factors:

i) *Selection bias* - patients who choose to get screened are more health conscious, better educated and have an inherently better prognosis. Selection bias can also occur when subjects decide to get screened because they have symptoms.

ii) *Lead-time* Screen-detected cases will survive longer even without benefit of early treatment, simply because they are detected earlier! This is shown in Figure 2 - the average survival duration is increased from 1.3 years (with no screening - see panel a.) to 2.5 years (with screening - see panel b.) purely due to the fact that the 3 subjects with disease were identified earlier (i.e., survival is increased due to the lead time).

Figure 2. Effect of screening and effective treatment on survival duration and mortality rate



a) No screening

1,000 population, followed for 5 years, 5 cases

Av. duration of survival= 6.5/5 = **1.3 yrs**

MR= 5/[(995X5)+5+4+3+2+2] x 100,000

= 5/ 4991 x 100,000 = **100/100,000 py's**

b) Screening with no reduction in MR

1,000 population, followed for 5 years, 5 cases

Av. duration of survival= 12.5/5 = **2.5 yrs**

MR= 5/[(995X5)+5+4+3+2+2] x 100,000

= **100/100,000 py's**

c) Screening with reduction in MR

1,000 population, followed for 5 years, 5 cases

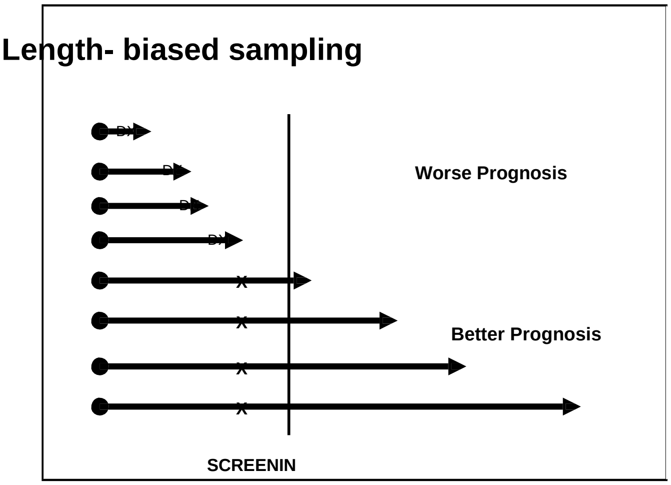
Av. duration of survival= 12.5/5 = **2.5 yrs**

MR= 4/[(995X5)+5+4+3+2+2] x 100,000

= 4/4991 x 100,000 = **80/100,000 py's**

iii) *Length-biased sampling* - screen detected cases are not simply a sample of *all* cases in a population but represent a sample of cases *prevalent in the asymptomatic (pre- clinical) phase*. Screening preferentially identifies slow growing, indolent cases that have a long pre-clinical phase. Slow growing tumours will obviously have a better prognosis because they have both a long pre-clinical phase and a long clinical phase, as illustrated in Figure 3.

Figure 3. Length-biased sampling (from Fletcher et al., 1997)



2) Disease-specific mortality rate (DSMR)
The *only* truly valid measure of the efficacy of a screening program is to conduct a randomized screening trial where the DSMR in the group assigned to screening is compared to the group assigned to no screening. Unlike the survival duration, the DSMR will not be changed by early diagnosis (i.e., lead time). This concept is illustrated in Figure 2. In panel c, screening has resulted in a mortality reduction after 5 years, since the first subject no longer dies at year 5, (and so continues to live as indicated by the arrow). The average survival duration calculated at the end of year 5 is still 2.5 years. However, since there are now only 4 deaths (as opposed to 5 deaths previously) the DSMR drops from 100 per 100,000 person years to 80 per 100,000 (equivalent to a 20% reduction in mortality). Thus, it is the DSMR and not the survival duration, that accurately reflects the benefit of screening.

There is one caveat about the DSMR however; within the confines of a screening trial the specific cause of death is usually assigned by an adjudication committee. Ideally this assignment is done without knowledge of the screening group that a particularly subject was assigned to. However, maintain blinding to this fact is often difficult, especially given that there is usually a detailed examination of the specific circumstances around each death. If the original random assignment becomes unblinded there is the real potential for bias to be introduced into the process. For example, in a breast cancer trial, there might be a tendency to call deaths that occurred in the mammography group not breast cancer related, while in the control group there might be a tendency to overdiagnose breast cancer as a cause of death. Because of these difficulties there is now considerable debate in the screening community that the ideal measure of screening efficacy should be all-cause mortality rather than the DSMR, because all-cause mortality is clearly not subject to these same biases (the subject is either dead or alive) (see Black WC, JNCI, 2002).

VI. Pseudo-disease or Over-diagnosis

One potential negative side-effect of screening is pseudo-disease or over-diagnosis which is the identification of disease that would not have become clinically apparent in the absence of screening. This can involve three forms:

i) Over-diagnosis - cases detected that would never have progressed to a clinical state – i.e., cancer cases with limited malignant potential. This is in fact an extreme form of length-biased sampling. A classic example is pap testing which despite reducing the incidence of invasive cervical cancer results in a large increase in the overall incidence of cervical cancer because of the “over-diagnosis’ of carcinoma in situ (See Table below). Other examples include PSA testing and low-grade prostate cancer, and mammography and ductal carcinoma in situ (see Ernster et al, 1996).

Table. Example of over-diagnosis: increase in carcinoma in situ of the cervix following introduction of mass screening (pap testing) in Connecticut (Laskey et al 1976)

	Age-adjusted Incidence rate (per 100,000)			
Year	Carcinoma in situ	Invasive carcinoma	Total	% in situ
1950-1954	3.8	18.1	21.9	17
1955-1959	9.7	17.1	26.8	36
1960-1964	18.8	13.6	32.4	58
1965-1969	28.6	11.6	40.2	71
1970-1973	32.8	10.9	43.7	75

ii) Competing risks - cases are identified that would have been interrupted by an unrelated death. An example would be the identification of prostate cancer in an 85 year old man who would have died of stroke.

iii) Serendipity - the identification of disease due to diagnostic testing that comes about for another reason. Example, chest x-ray for TB screening that identifies lung cancer, or colonoscopy detection of colorectal cancer following a positive FOBT (Ederer et al, 1997).

VII. Feasibility and Need for Screening

There are several other important issues beyond the demonstration that screening leads to decreased mortality and/or morbidity that need to be addressed before deciding to invoke a screening program. These include:

a) Acceptability: The program should be convenient, free of discomfort, efficient and economical.

b) Efficiency: A low PVP indicates a wasteful program, as most of the test positive individuals worked up will not have disease.
A high PVP can still be associated with only a few cases detected and a small reduction in overall mortality.
If mortality from the disease is normally low or if the risk of death from other causes is high (for example in the very aged) then the screening program will not reduce mortality very much.

c) Cost-effectiveness: Should these health care dollars be spent on this program? Most population based screening programs run about $30-50,000 per year of life saved (or higher).

Another way of evaluating the need and feasibility of screening is to place all the subjects who would develop the condition you are trying to help (e.g., lung cancer or prostate cancer) into one of the following three groups:

1. A cure is necessary but not possible (Nec,NotPos). In other words, if the target condition is death from lung cancer, these subjects are going to die of lung cancer regardless, and so would not be helped by a screening program.

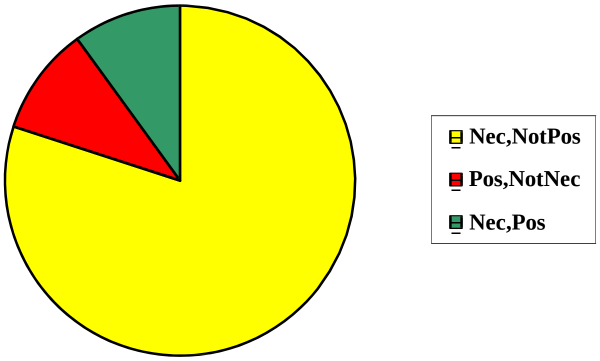
2. Cure is possible but not necessary (Pos,NotNec). This group includes subjects who develop lung cancer but will not die of it – this is an example of over-diagnosis (cases die of something else before dying of lung cancer). Again, a screening program will not be helpful to this group.

3. Cure is necessary and maybe possible (Nec,Pos) This is the only group that can benefit from screening! They represent cases of lung cancer who would have died of the disease had it not been for the effect of the screening program (this of course assumes that the screening program is effective in reducing the risk of death).

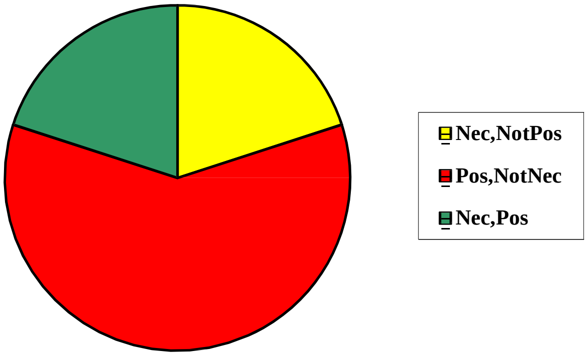
In terms of feasibility of screening it is helpful to consider the relative sizes of these three groups. While it is not possible to be absolutely sure of the sizes of these three groups, a reasonable estimate can be made based on knowledge of the natural history of disease, the potential of the intervention to identify the condition early, and potential effect of treatment to impact the outcome, and finally the potential to identify undiagnosed but benign disease.

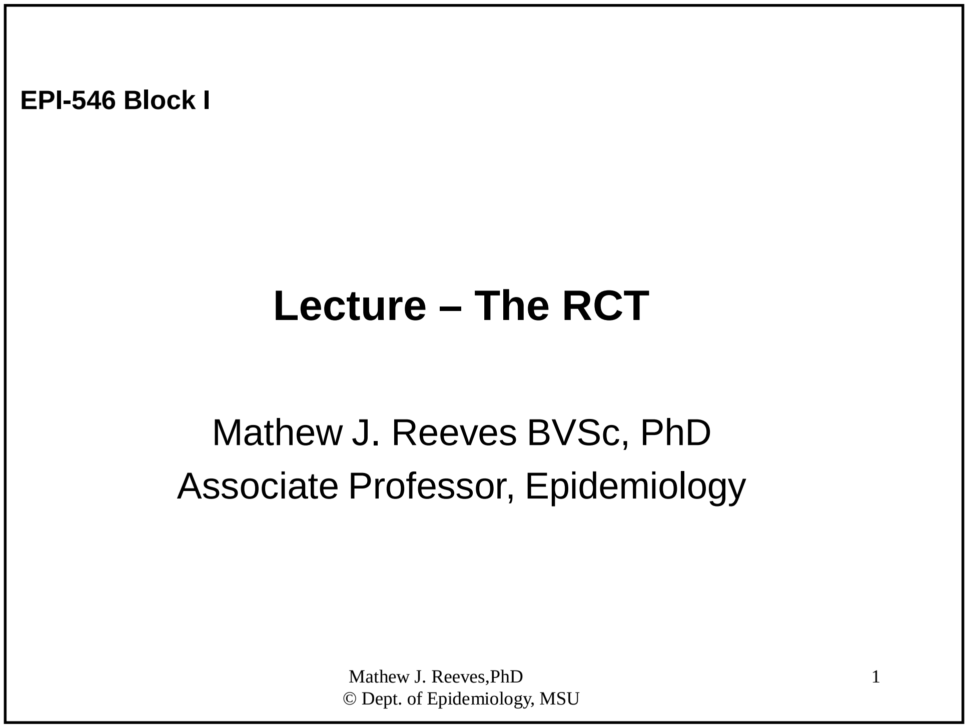
The charts below give a hypothetical example of this sort of assessment for two cancers. For Lung Cancer, the size of group 3 (Nec,Pos) is maybe 10%, but 80% are in group 1 (Nec, Not Pos) and hence can't be helped at all, while a further 10% are in group 2 (Pos, Not Nec) and don't need to be helped. For prostate cancer, the size of group 3 (Nec,Pos) is maybe 20%, but now we have a lot more men, say 60%, who are in group 2 (Pos, Not Nec) – these represent men with low grade or "benign" acting prostate cancer that will not kill them. Finally, there is another 20% in group 1 (Nec, Not Pos) who will die of prostate cancer regardless of any screening program. Obviously you would like most people to be in group 3 because this represent the positive (worthwhile) effects of screening. You would also like to keep group 2 as small as possible, since this represent the negative (or wasteful) effects of a screening program.

Lung Cancer:



Prostate Cancer:

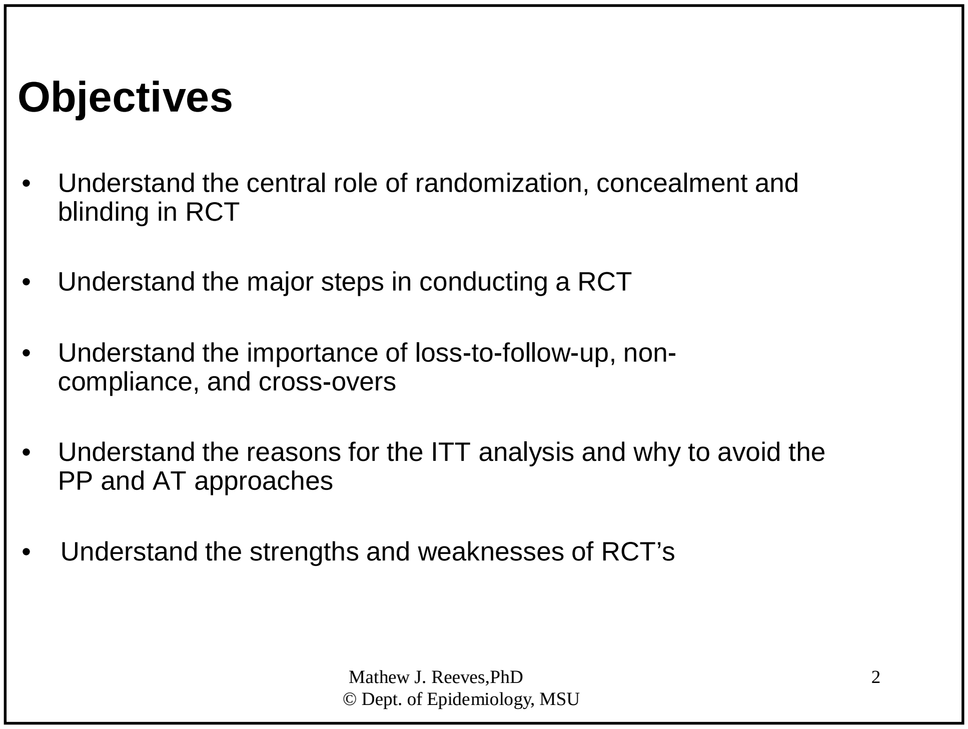




EPI-546 Block I

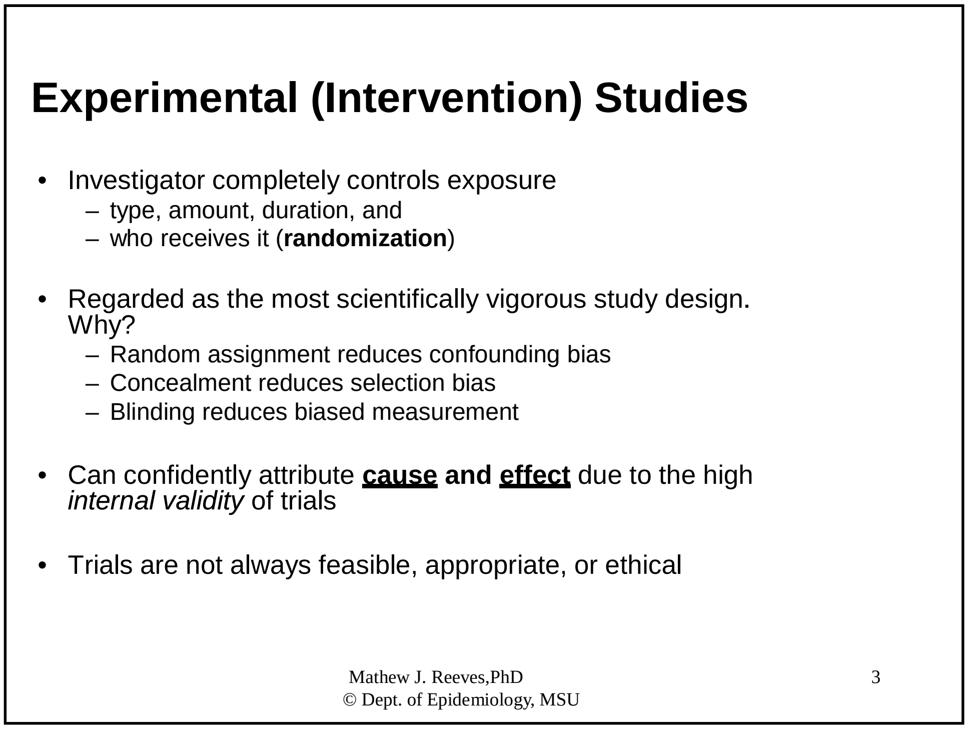
Lecture – The RCT

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology



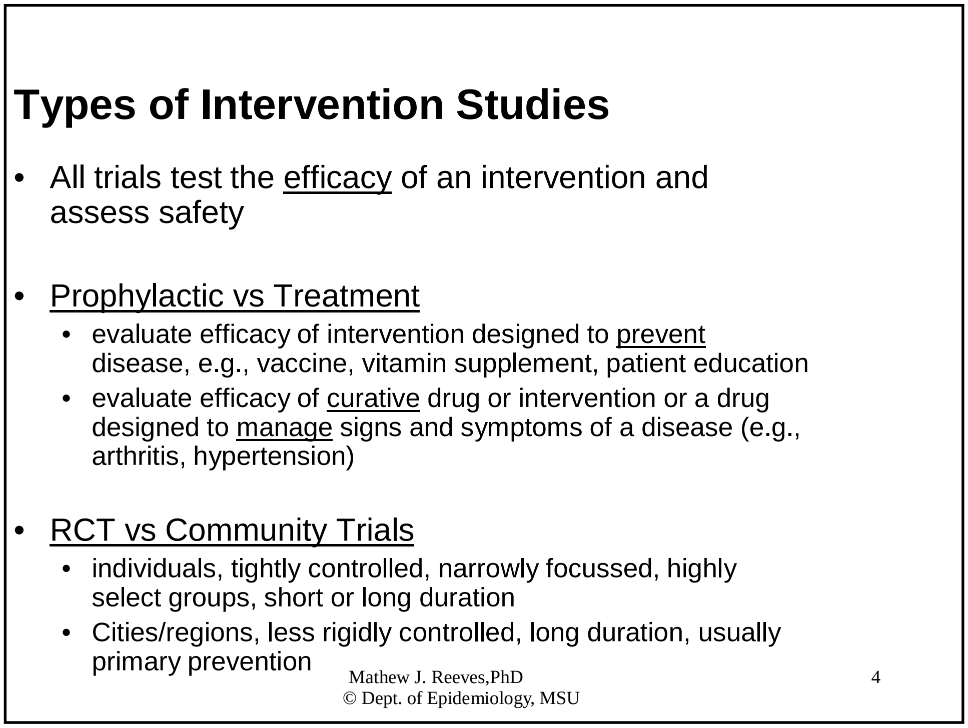
Objectives

- Understand the central role of randomization, concealment and blinding in RCT
- Understand the major steps in conducting a RCT
- Understand the importance of loss-to-follow-up, non-compliance, and cross-overs
- Understand the reasons for the ITT analysis and why to avoid the PP and AT approaches
- Understand the strengths and weaknesses of RCT’s



Experimental (Intervention) Studies

- Investigator completely controls exposure
  - type, amount, duration, and
  - who receives it (**randomization**)
- Regarded as the most scientifically vigorous study design. Why?
  - Random assignment reduces confounding bias
  - Concealment reduces selection bias
  - Blinding reduces biased measurement
- Can confidently attribute **<u>cause</u> and <u>effect</u>** due to the high *internal validity* of trials
- Trials are not always feasible, appropriate, or ethical



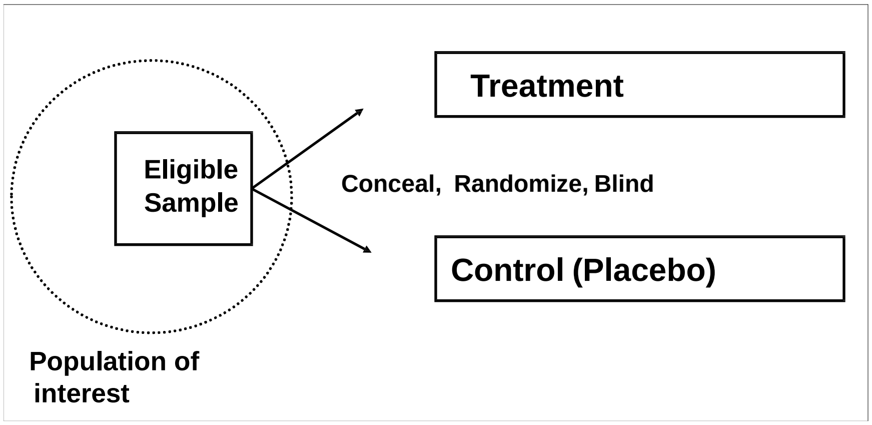
Types of Intervention Studies

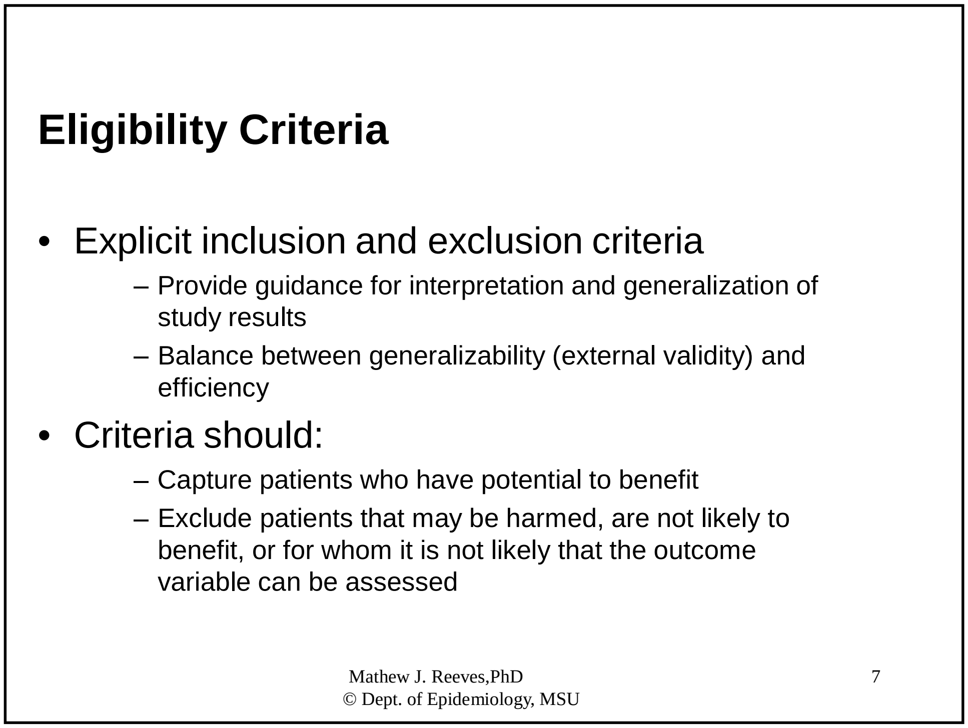
- All trials test the <u>efficacy</u> of an intervention and assess safety
- <u>Prophylactic vs Treatment</u>
  - evaluate efficacy of intervention designed to <u>prevent</u> disease, e.g., vaccine, vitamin supplement, patient education
  - evaluate efficacy of <u>curative</u> drug or intervention or a drug designed to <u>manage</u> signs and symptoms of a disease (e.g., arthritis, hypertension)
- <u>RCT vs Community Trials</u>
  - individuals, tightly controlled, narrowly focussed, highly select groups, short or long duration
  - Cities/regions, less rigidly controlled, long duration, usually primary prevention

RCT’s – Overview of the Process

- 1. Inclusion Criteria
- 2. Exclusion Criteria
- 3. Baseline Measurements
- 4. Randomization and concealment
- 5. Intervention
- 6. Blinding
- 7. Follow-up (FU) and Compliance
- 8. Measuring Outcomes
- 9. Statistical Analyses

RCT: Basic Design

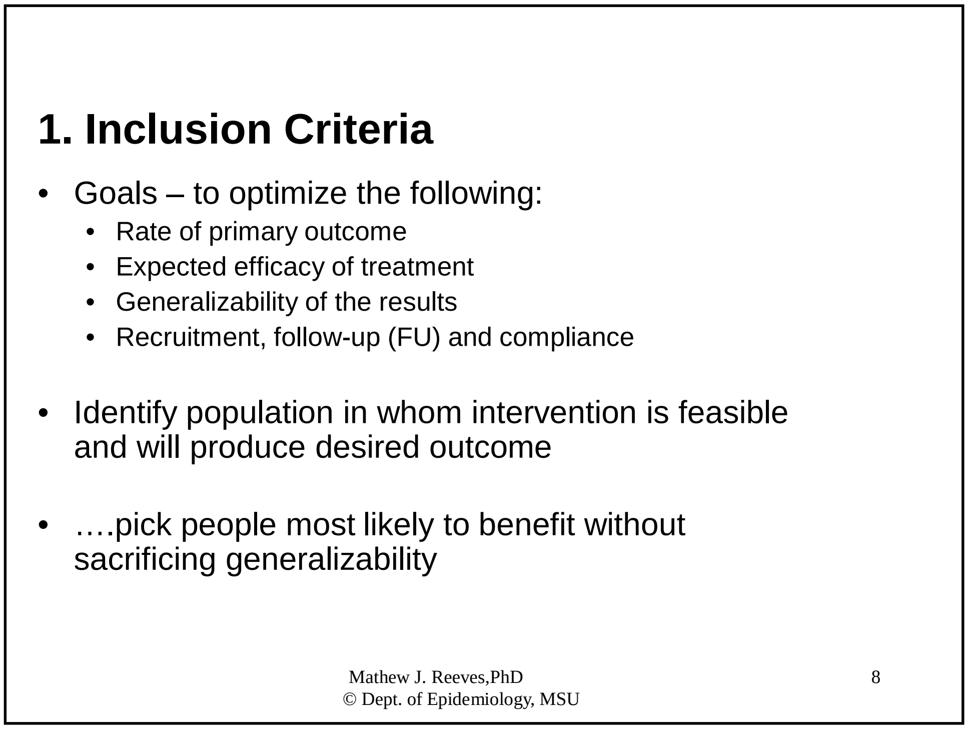




Eligibility Criteria

- Explicit inclusion and exclusion criteria
  - – Provide guidance for interpretation and generalization of study results
  - – Balance between generalizability (external validity) and efficiency
- Criteria should:
  - – Capture patients who have potential to benefit
  - – Exclude patients that may be harmed, are not likely to benefit, or for whom it is not likely that the outcome variable can be assessed

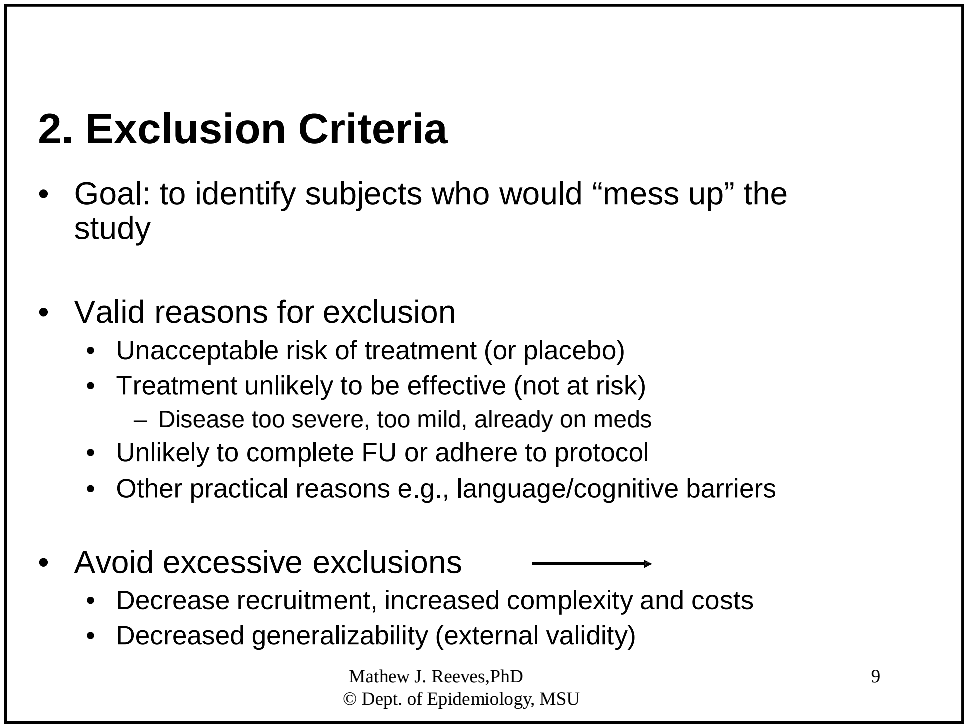
Mathew J. Reeves,PhD
© Dept. of Epidemiology, MSU
7



1. Inclusion Criteria

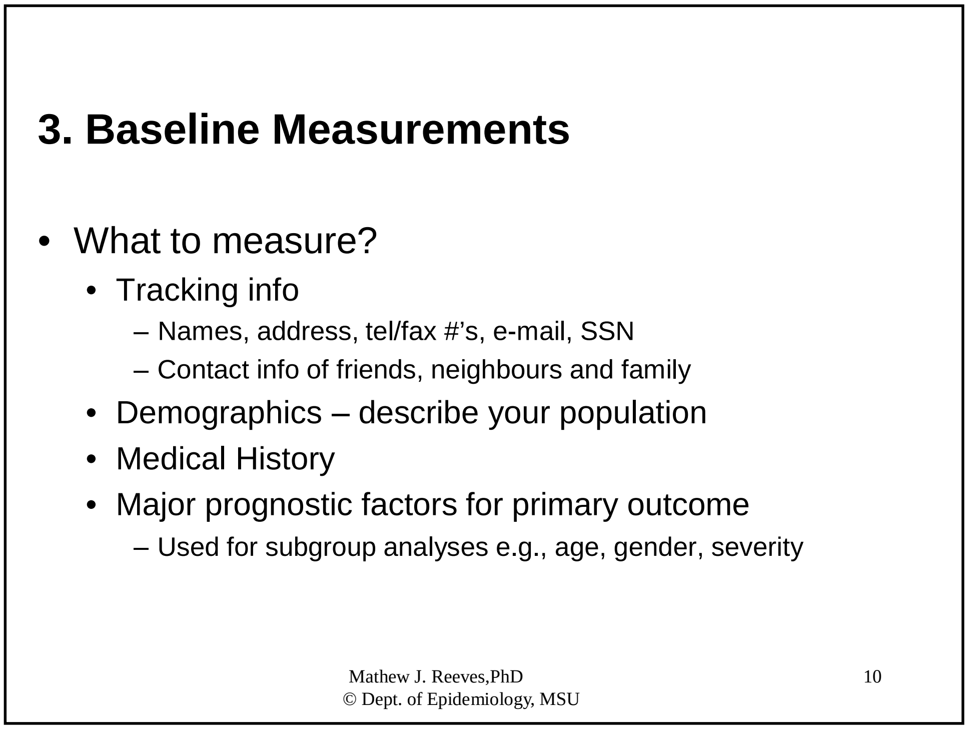
- Goals – to optimize the following:
  - Rate of primary outcome
  - Expected efficacy of treatment
  - Generalizability of the results
  - Recruitment, follow-up (FU) and compliance
- Identify population in whom intervention is feasible and will produce desired outcome
- ….pick people most likely to benefit without sacrificing generalizability

Mathew J. Reeves,PhD
© Dept. of Epidemiology, MSU
8



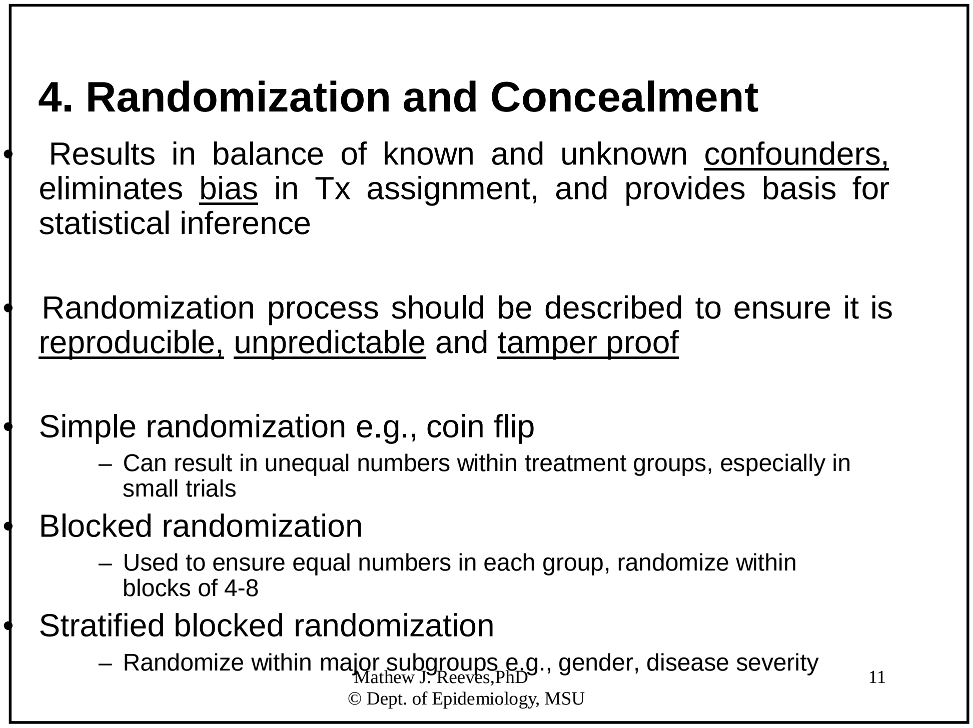
2. Exclusion Criteria

- Goal: to identify subjects who would “mess up” the study
- Valid reasons for exclusion
  - Unacceptable risk of treatment (or placebo)
  - Treatment unlikely to be effective (not at risk)
    - Disease too severe, too mild, already on meds
  - Unlikely to complete FU or adhere to protocol
  - Other practical reasons e.g., language/cognitive barriers
- Avoid excessive exclusions →
  - Decrease recruitment, increased complexity and costs
  - Decreased generalizability (external validity)



3. Baseline Measurements

- What to measure?
  - Tracking info
    - Names, address, tel/fax #’s, e-mail, SSN
    - Contact info of friends, neighbours and family
  - Demographics – describe your population
  - Medical History
  - Major prognostic factors for primary outcome
    - Used for subgroup analyses e.g., age, gender, severity

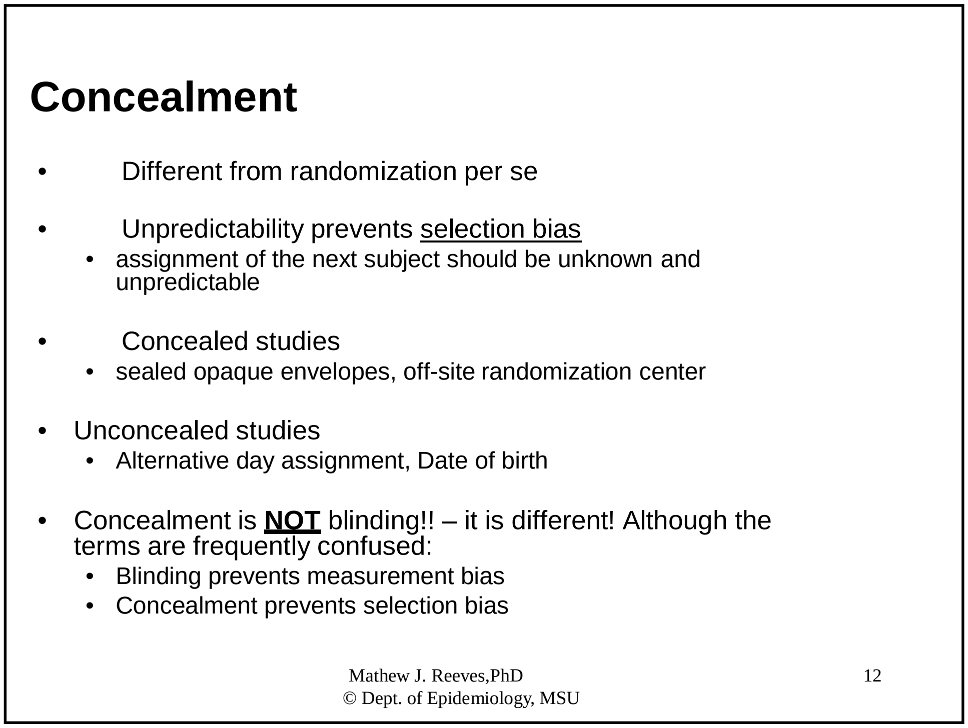


4. Randomization and Concealment

- Results in balance of known and unknown <u>confounders,</u> eliminates <u>bias</u> in Tx assignment, and provides basis for statistical inference

- Randomization process should be described to ensure it is <u>reproducible,</u> <u>unpredictable</u> and <u>tamper proof</u>

- Simple randomization e.g., coin flip
  – Can result in unequal numbers within treatment groups, especially in small trials
- Blocked randomization
  – Used to ensure equal numbers in each group, randomize within blocks of 4-8
- Stratified blocked randomization
  – Randomize within major subgroups e.g., gender, disease severity



Concealment

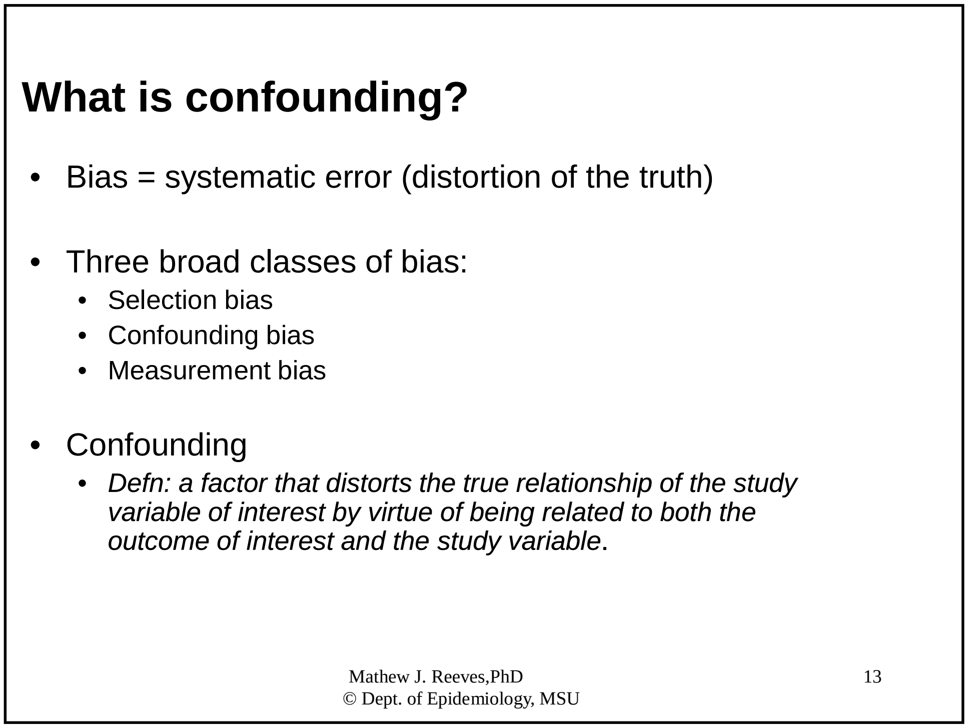
- Different from randomization per se

- Unpredictability prevents <u>selection bias</u>
  - assignment of the next subject should be unknown and unpredictable

- Concealed studies
  - sealed opaque envelopes, off-site randomization center

- Unconcealed studies
  - Alternative day assignment, Date of birth

- Concealment is **<u>NOT</u>** blinding!! – it is different! Although the terms are frequently confused:
  - Blinding prevents measurement bias
  - Concealment prevents selection bias

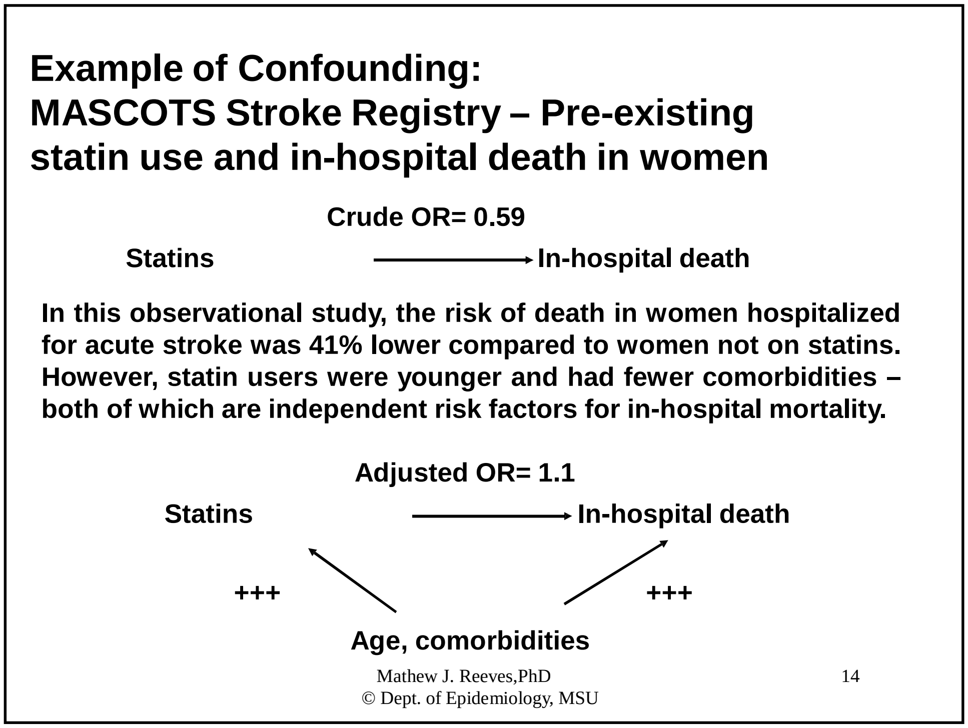


What is confounding?

- Bias = systematic error (distortion of the truth)

- Three broad classes of bias:
  - Selection bias
  - Confounding bias
  - Measurement bias

- Confounding
  - *Defn: a factor that distorts the true relationship of the study variable of interest by virtue of being related to both the outcome of interest and the study variable*.



Example of Confounding: MASCOTS Stroke Registry – Pre-existing statin use and in-hospital death in women

Crude OR= 0.59

Statins → **In-hospital death**

In this observational study, the risk of death in women hospitalized for acute stroke was 41% lower compared to women not on statins. However, statin users were younger and had fewer comorbidities – both of which are independent risk factors for in-hospital mortality.

Adjusted OR= 1.1

Statins → **In-hospital death**

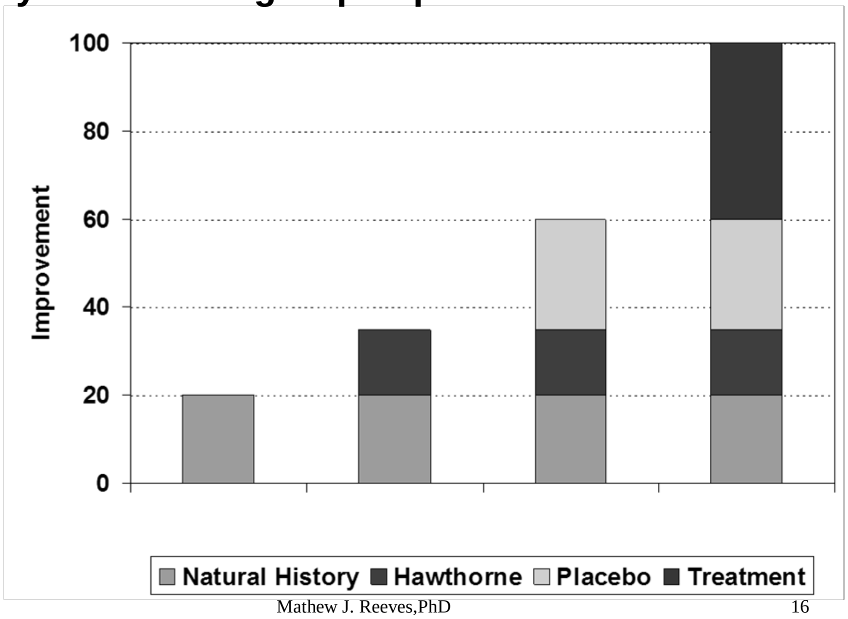
+++ **+++**

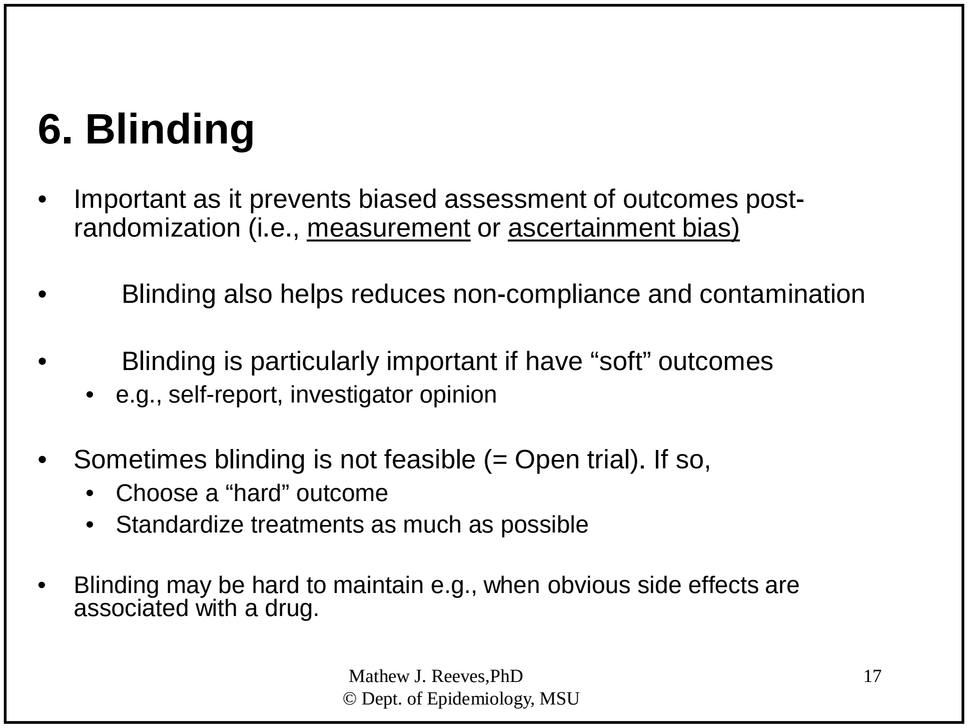
Age, comorbidities

5. Intervention

- Balance between <u>efficacy</u> & <u>safety</u>
  - Everyone is exposed to potential side effects but only the few who develop dis/outcome can benefit
  - Hence, usually use “lowest effective dose”
  - Most trials are under-powered to detect side effects – hence Phase IV trials and post-marketing surveillance
- Control group:
  - Placebo or standard treatment
  - Risk of contamination (getting the treatment somewhere else)

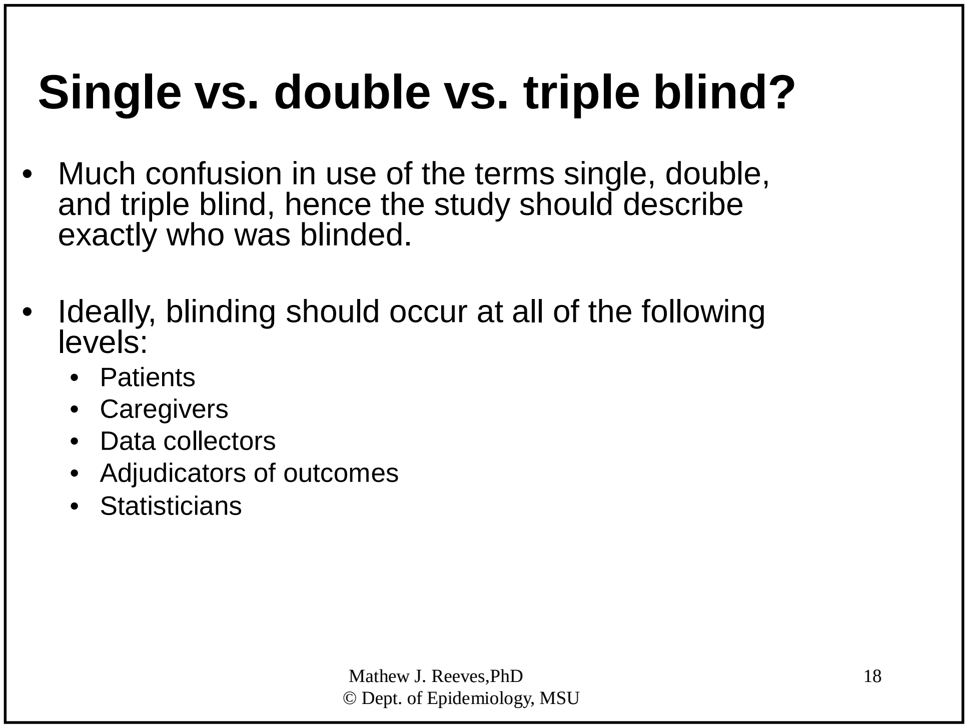
Why is a control group important?





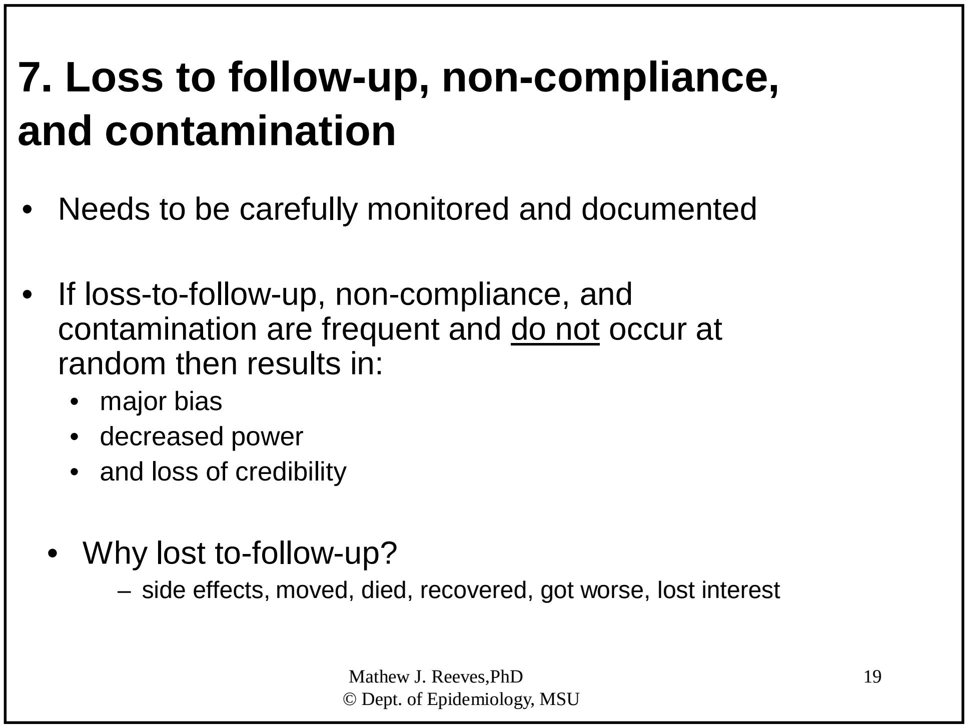
6. Blinding

- Important as it prevents biased assessment of outcomes post-randomization (i.e., measurement or ascertainment bias)
- Blinding also helps reduces non-compliance and contamination
- Blinding is particularly important if have “soft” outcomes
  - e.g., self-report, investigator opinion
- Sometimes blinding is not feasible (= Open trial). If so,
  - Choose a “hard” outcome
  - Standardize treatments as much as possible
- Blinding may be hard to maintain e.g., when obvious side effects are associated with a drug.



Single vs. double vs. triple blind?

- Much confusion in use of the terms single, double, and triple blind, hence the study should describe exactly who was blinded.
- Ideally, blinding should occur at all of the following levels:
  - Patients
  - Caregivers
  - Data collectors
  - Adjudicators of outcomes
  - Statisticians

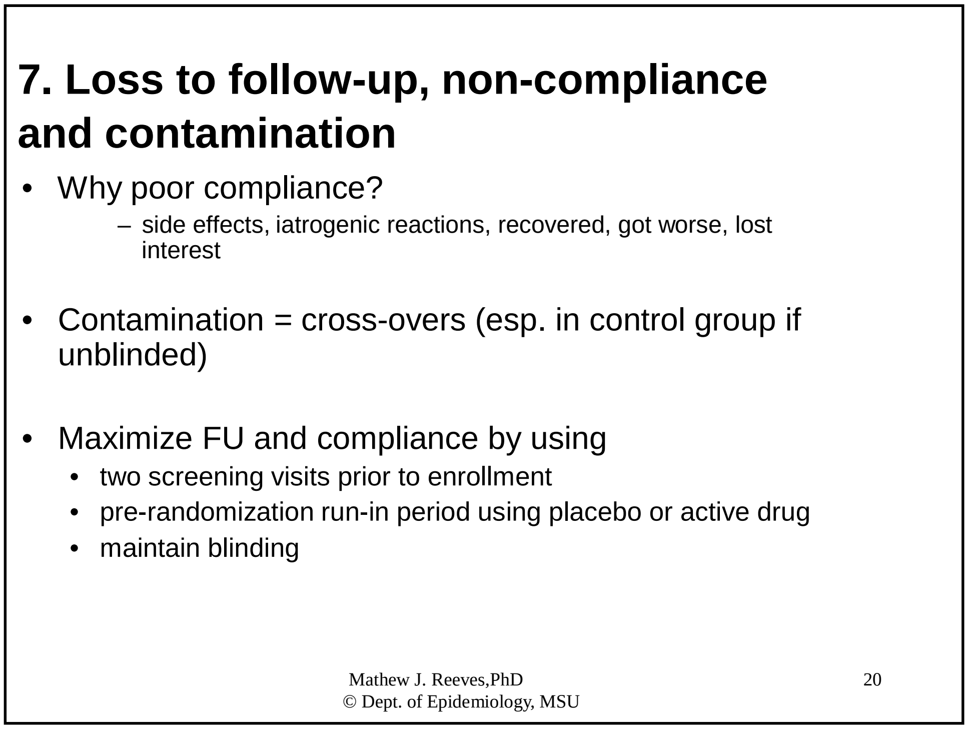


7. Loss to follow-up, non-compliance, and contamination

- Needs to be carefully monitored and documented

- If loss-to-follow-up, non-compliance, and contamination are frequent and do not occur at random then results in:
  - major bias
  - decreased power
  - and loss of credibility

- Why lost to-follow-up?
  - side effects, moved, died, recovered, got worse, lost interest



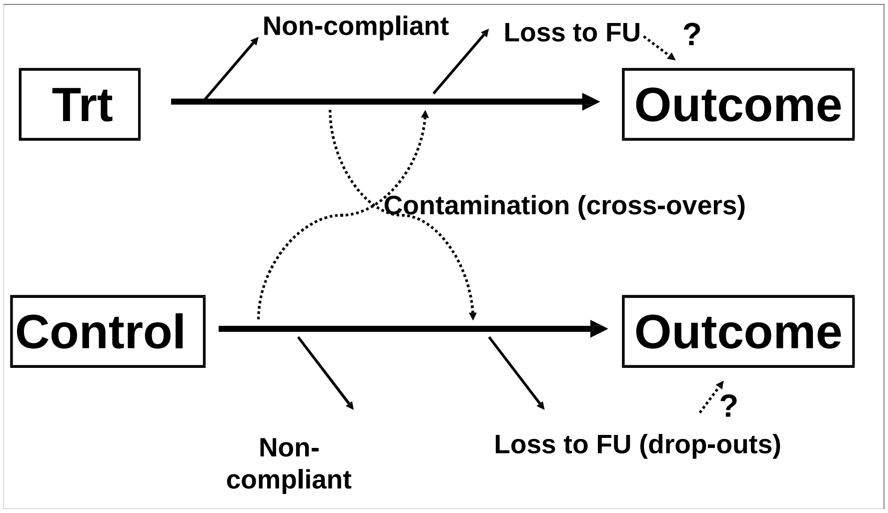
7. Loss to follow-up, non-compliance and contamination

- Why poor compliance?
  - side effects, iatrogenic reactions, recovered, got worse, lost interest

- Contamination = cross-overs (esp. in control group if unblinded)

- Maximize FU and compliance by using
  - two screening visits prior to enrollment
  - pre-randomization run-in period using placebo or active drug
  - maintain blinding

Loss to FU, poor compliance, contamination = Slow death of a trial



8. Measuring Outcomes

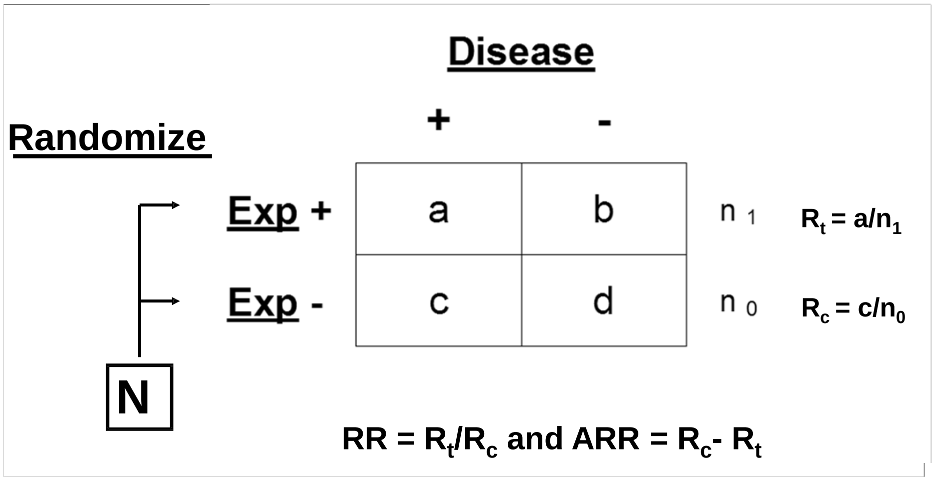
- Best = Hard clinically relevant end points
  - e.g., disease rates, death, recovery, complications
  - Must be measured with accuracy and precision
  - Must be monitored equally in both groups
- Surrogate end points
  - used in short-term clinical trials or to reduce size/length of follow-up
    - – Must be biologically plausible
    - – Association between surrogate and a hard endpoint must have been previously demonstrated
    - – e.g., reduced BP in lieu of stroke incidence
- Best to use blinded adjudication, especially for soft outcomes

9. Statistical Analyses

- On the surface analysis is relatively simple because of RCT design:
  - Measure clinical effect with 95% CI using
    - RR, RRR, ARR, NNT
  - For time-dependent outcomes use Kaplan-Meire curves, and/or Cox regression modeling
  - Test for statistical significance (p-value):
    - Categorical outcome – Chi-square test
    - Continuous outcome – t-test
    - Or use non-parametric methods where necessary

RCT's – Basic Analysis

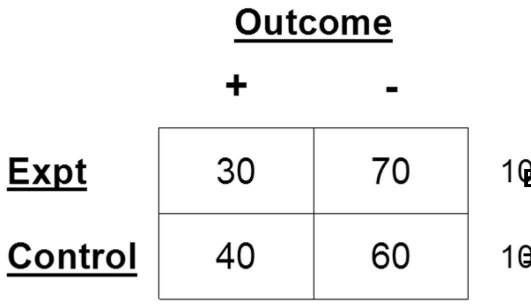
Dichotomous (Disease Yes/No) Outcome



	Disease +	**Disease** -		
Exp +	a	b	n_1	$R_t = a/n_1$
Exp -	c	d	n_0	$R_c = c/n_0$

$RR = R_t/R_c$ and $ARR = R_c - R_t$

Measures of Effect used in RCT's



	Outcome		
	+	-	
Expt	30	70	100
Control	40	60	100

EER = 30/100= 30% CER

= 40/100 = 40%

EER = Experimental (or treatment) event rate
CER = Control (or baseline) event rate

RR = EER/CER = 30/40 = 75%
RRR = 1 – RR = 25%
ARR = CER - EER = 40 – 30 = 10%
NNT = 1/ARR = 1/0.10 = 10

9. Statistical Analyses - ITT

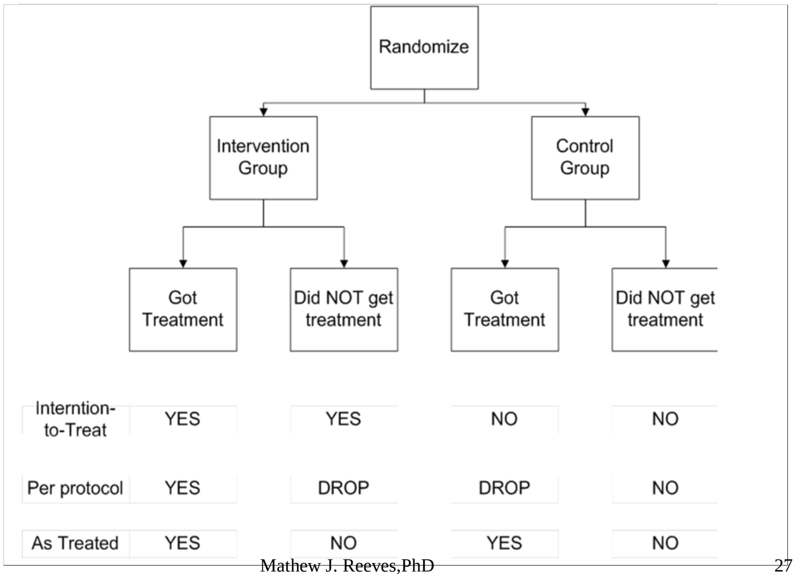
- Intention-to-Treat Analysis
  - Gold Standard
  - Compares outcomes based on original randomization scheme regardless of eligibility, non-compliance, cross-overs, and lost-to-follow-up

- Per Protocol (PP) Analysis
  - Compares outcomes based on actual treatment received among those who were compliant (analysis drops non-compliant)
  - Asks whether the treatment works among only those that comply

- As Treated (AT) Analysis
  - Compares outcomes based on actual treatment received regardless of original assignment.
  - Equivalent to analyzing the data as a cohort study!
  - Asks whether the treatment works among those that took it.

- Both PP and AT approaches ignore original randomization and are therefore subject to BIAS!!!

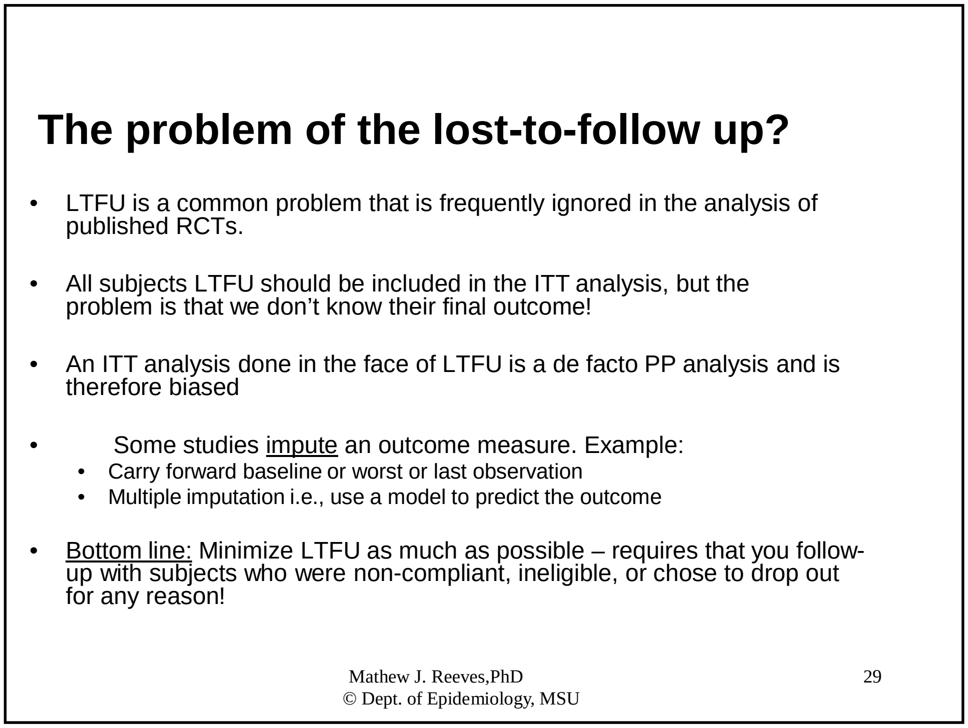
ITT vs. Per-Protocol vs. As Treated



	Got Treatment	Did NOT get treatment	Got Treatment	Did NOT get treatment
Interntion-to-Treat	YES	YES	NO	NO
Per protocol	YES	DROP	DROP	NO
As Treated	YES	NO	YES	NO

Effect of ITT vs. PP vs. AT analyses on an RCT of coronary artery bypass surgery versus medical treatment in 767 men with stable angina. (Lancet 1979;i:889-93).

	Allocated (vs. actual) treatment				
	Medical (medical)	Medical (surgical)	Surgical (surgical)	Surgical (medical)	ARR (95% CI)
Subjects	323	50	368	26	
Deaths	27	2	15	6	
Mortality	8.4%	4.0%	4.1%	23.1%	
ITT analysis	7.8% (29/373)		5.3% (21/394)		2.4% (-1.0, 6.1)
PP analysis	8.4% (27/323)		4.1% (15/368)		4.3% (0.7, 8.2)
AT analysis	9.5% (33/349)		4.1% (17/418)		5.4% (1.9, 9.3)



The problem of the lost-to-follow up?

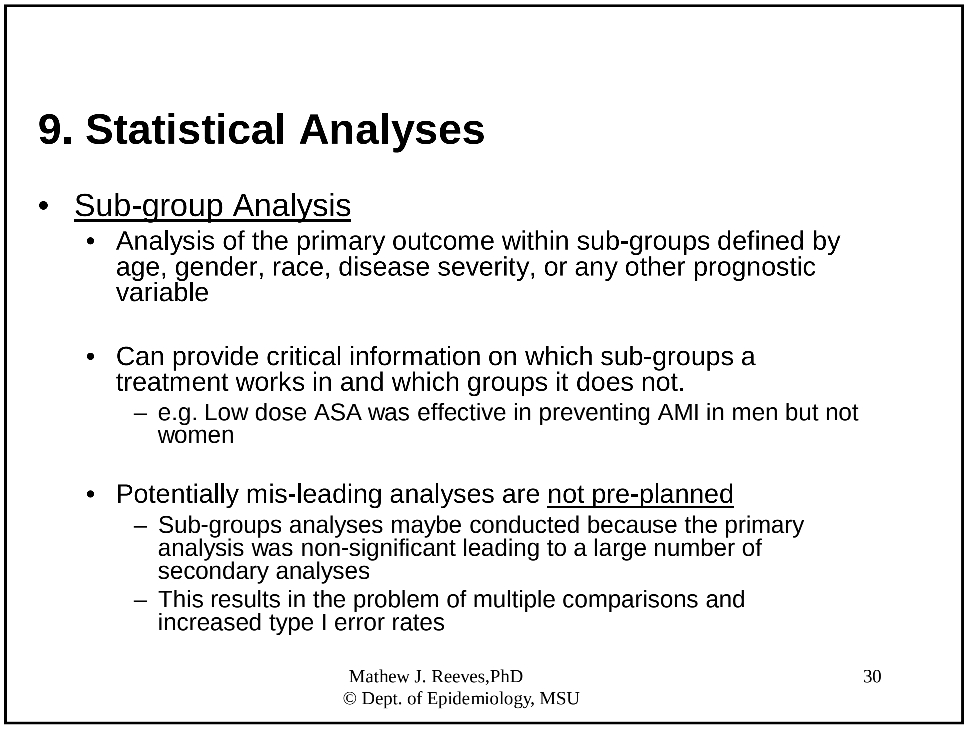
- LTFU is a common problem that is frequently ignored in the analysis of published RCTs.

- All subjects LTFU should be included in the ITT analysis, but the problem is that we don’t know their final outcome!

- An ITT analysis done in the face of LTFU is a de facto PP analysis and is therefore biased

- Some studies impute an outcome measure. Example:
  - Carry forward baseline or worst or last observation
  - Multiple imputation i.e., use a model to predict the outcome

- Bottom line: Minimize LTFU as much as possible – requires that you follow-up with subjects who were non-compliant, ineligible, or chose to drop out for any reason!



9. Statistical Analyses

- Sub-group Analysis
  - Analysis of the primary outcome within sub-groups defined by age, gender, race, disease severity, or any other prognostic variable

  - Can provide critical information on which sub-groups a treatment works in and which groups it does not.
    - e.g. Low dose ASA was effective in preventing AMI in men but not women

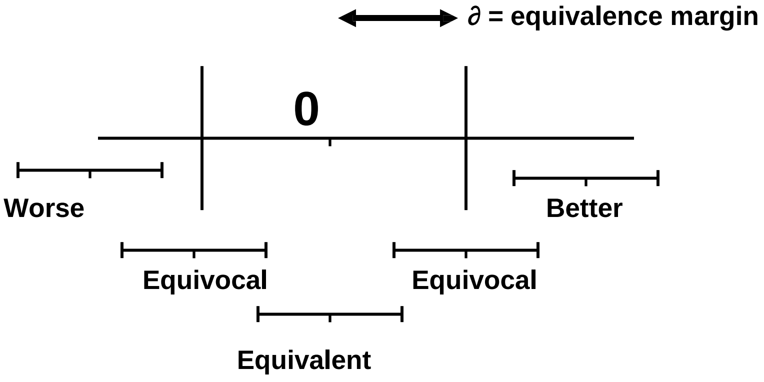
  - Potentially mis-leading analyses are not pre-planned
    - Sub-groups analyses maybe conducted because the primary analysis was non-significant leading to a large number of secondary analyses
    - This results in the problem of multiple comparisons and increased type I error rates

10. Equivalence and Non-inferiority Trials

- Equivalence trials
  - A trial designed to prove that a new drug is equivalent to an existing standard drug with a given tolerance (∂) or equivalence margin
  - Most often used in the area of generic drug development to prove that the new generic drug is bio-equivalent to the original drug
    - i.e., similar bioavailability, pharmacology etc

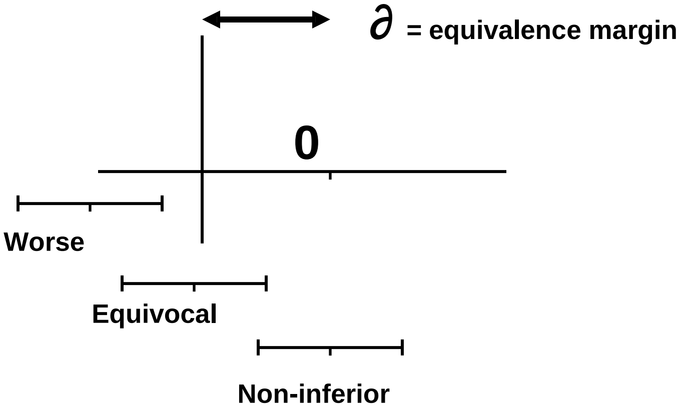
- Non-inferiority trials
  - A trial designed to prove that a new drug is no less effective than an existing standard drug
    - This is a one-sided equivalence test
    - More common especially as it is becoming more difficult to prove superiority of newer treatments
  - Example
    - GUSTO III, NEJM 1997 (reteplase vs.alteplase for treatment of AMI)

Equivalence Trials



95% CIs

Non-inferiority Trials



∂ = equivalence margin

0

Worse

Equivocal

Non-inferior

Summary of RCTs

- Advantages
  - High internal validity
  - Able to control selection, confounding and measurement biases
  - True measure of treatment efficacy (cause and effect)
- Disadvantages
  - Low external validity (generalizability)
  - Strict enrollment criteria creates a unique, highly selected study population
  - Complicated, expensive, time consuming
  - Ethical and practical limitations

EPI-546: Fundamentals of Epidemiology and Biostatistics

Course Notes - The RCT

Mat Reeves BVSc, PhD

God gave us randomization so we can detect the modest effects of most treatments

Objectives

1. Understand, explain, and distinguish between randomization, concealment and blinding.
2. Understand and explain how randomization, concealment and blinding prevent confounding, selection, and measurement biases, respectively.
3. Understand the difference between internal and external validity in the context of a RCT.
4. Understand the importance of the target population and how this is influenced by inclusion and exclusion criteria.
5. Understand the major sequential steps in designing and conducting a RCT.
6. Understand, explain, and distinguish between loss-to-follow-up, non-compliance, and cross-overs.
7. Understand, explain and differentiate between the intention-to-treat (ITT), per-protocol (PP) and as-treated (AT) analyses, and understand the role of non-compliance and cross-overs.
8. Understand the importance and impact of loss-to-follow on the ITT analysis (need for imputation) and determine the impact of LTFU by best-case, worst–case analysis.
9. Outcome measures – distinguish between composite vs. individual measures, patient orientated vs. surrogate outcomes, pre-defined vs. post-hoc outcomes.
10. Understand the basic organization of a RCT publication (Flow diagram of patient enrollment and follow-up, Table 1 comparisons of baseline characteristics).
11. Describe the advantages and disadvantages of trials.

Outline:

I. Introduction to the RCT
II. An Overview of the RCT Design
   1. Inclusion Criteria
   2. Exclusion Criteria
   3. Baseline Measurements
   4. Randomization and concealment
   5. Intervention
   6. Blinding
   7. Follow-up (FU), Non-Compliance and Contamination
   8. Measuring Outcomes, Sub-group analyses and Surrogate end points
   9. Statistical Analyses (ITT vs. PP vs. AS)
   10. RCTs and Meta-analyses - assessing trial quality, reporting, and trial registration
   11. Equivalence and Non-inferiority Designs

III. Advantages and disadvantages of RCT’s

I. Introduction to the RCT

The randomized clinical trial is an *experimental study* conducted on clinical patients (with their consent of course!). The investigator seeks to completely control the exposure (in terms of its type, amount, and duration), and (most importantly) who receives it through the process of randomization.

RCT's are regarded as the most scientifically vigorous study design. Because:

- An unpredictable (i.e., concealed) random assignment eliminates (or at least greatly reduces) confounding from known and unknown prognostic factors (that is, it makes the groups equivalent in terms of their prognosis at baseline).

- Blinding eliminates biased measurement, so that outcomes are measured with the same degree of accuracy and completeness in every participant.

Because of these conditions, it is then possible to confidently attribute **cause and effect** – that is, because the only thing that differed between the groups (or arms) of the trial was the presence or absence of the intervention, any effect can be ascribed to it (assuming a well conducted, unbiased study). The RCT is therefore described as having high *internal validity* – the experimental design ensures that, within reason, strong cause and effect conclusions can be drawn from the results. While RCT's are the gold standard by which we determine the efficacy of treatments, trials are not always feasible, appropriate, or ethical.

There are two types of RCTs:

Prophylactic trials

- evaluate the efficacy of an intervention designed to prevent disease, e.g., vaccine, vitamin supplement, patient education, screening.

Treatment trials

- evaluate efficacy of a curative drug or intervention or a drug or intervention designed to manage or mitigate signs and symptoms of a disease (e.g., arthritis, hypertension).

Also, RCTs can be done at the individual level, where highly select groups of individuals are randomized under tightly controlled conditions, or they can be done at the community level, where large groups are randomized (e.g., cities, regions) under less rigidly controlled conditions. Community trials are usually conducted to test interventions for primary prevention purposes – so they are prophylactic trials done on a wide scale.

II An Overview of the RCT Design and Process

1. Inclusion Criteria

Specific inclusion criteria are used to optimize the following:

- The rate of the primary outcome
- The expected efficacy of treatment

- The generalizability of the results
- The recruitment, follow-up, and compliance of patients

So, the goal is to identify the sub-population of patients in whom the intervention is feasible, and that will produce the desired effect. To this end, the choice of inclusion criteria represents a balance between picking the people who are most likely to benefit without sacrificing the generalizability of the study. If you make the inclusion criteria too restrictive you run the risk of having the study population so unique that no one else will be able to apply your findings to their population.

2. Exclusion Criteria

Specific exclusion criteria are used to exclude subjects who would "mess up" the study. Valid reasons for excluding patients might include the following:

- When the risk of treatment (or placebo) is unacceptable.
- When the treatment is unlikely to be effective because the disease is too severe, or too mild, or perhaps the patient has already received (and failed) the treatment.
- When the patient has other conditions (co-morbidities) that would either interfere with the intervention, or the measurement of the outcome, or the expected length of follow-up e.g., terminal cancer.
- When the patient is unlikely to complete follow-up or adhere to the protocol.
- Other practical reasons e.g., language/cognitive barriers/no phone at home etc.

Again, one needs to be careful to avoid using excessive exclusions even when they appear perfectly rational. Excessive number of exclusions can add to the complexity to the screening process (remember, every exclusion criteria needs to be assessed in every patient), and ultimately to decreased recruitment. Once again, the choice of exclusion criteria represents a balance between picking subjects who are more likely to make your study a success without sacrificing *generalizability*. The real danger of setting very restrictive inclusion and exclusion criteria is that the final study population becomes so highly selective that nobody is interested in the results because they don't apply to real-world patients. In other words, while *internal validity* may have been maximized, the study's generalizability or *external validity* is badly compromised.

3. Baseline Measurements

At baseline you want to collect information that will:

- Describe the characteristics of the subjects in your study i.e., demographics. This is especially important in demonstrating that the randomization process worked (i.e., that you achieved balance between the intervention and control groups).
- Assist in the tracking the subject during the study (to prevent loss-to-follow-up). This includes names, address, tel/fax #'s, e-mail, SSN, plus contact information of friends, neighbours, and family (you can never collect too much information here).
- Identify major clinical characteristics and prognostic factors for the primary outcome that can be evaluated in pre-specified subgroup analyses. For example, if you thought that the treatment effect could be different in women compared to men, then you would collect information on gender, so that the effect could be evaluated within each of these two sub-groups.

The amount of baseline data that needs to be collected does not have to be excessive, because the randomization process should result in identical groups. However, it is always necessary to collect sufficient baseline information regarding demographics and clinical characteristics to prove this point.

4. Randomization and concealment

Randomization results in balance of known and unknown confounders. The randomization process should be reproducible, unpredictable, and tamper proof. It is critical that the randomization scheme itself should be **unpredictable** – so that it is not possible ahead of time to predict which group a given subject would be randomized to. Unpredictability is assured through the process of concealment which is critical in preventing selection bias – that is, the potential for investigators to manipulate who gets what treatment. Such "manipulation" in clinical trials has been well documented (for an example, see Schulz KF Subverting randomization in clinical trials. JAMA 1995;274:1456-8). Finally, note that the issues of randomization and concealment should be kept separate from blinding – they are completely different![1]

The actual process of generating the randomization scheme and the steps taken to ensure concealment should be described in detail - whether it is a simple coin toss (the least preferable method), or the use of sealed opaque envelopes, or a sophisticated off-site centralized randomization centre.

There are several different modifications to simple (individual level) randomization schemes such as a coin flip. Blocked randomization refers to the randomization done within blocks of 4 to 8 subjects. It is used to ensure that there is equal balance in the number of treatment and control subjects throughout the study.

Stratified blocked randomization refers to the process whereby strata are defined according to a critically important factor or subgroup (e.g., gender, disease severity, or study center), and the randomization process conducted within each strata. Again, this design ensures that there is balance in the number of treatment and control subjects within each of the sub-groups (so that at the end of the trial the factor in question is distributed equally among the treatment and control groups).

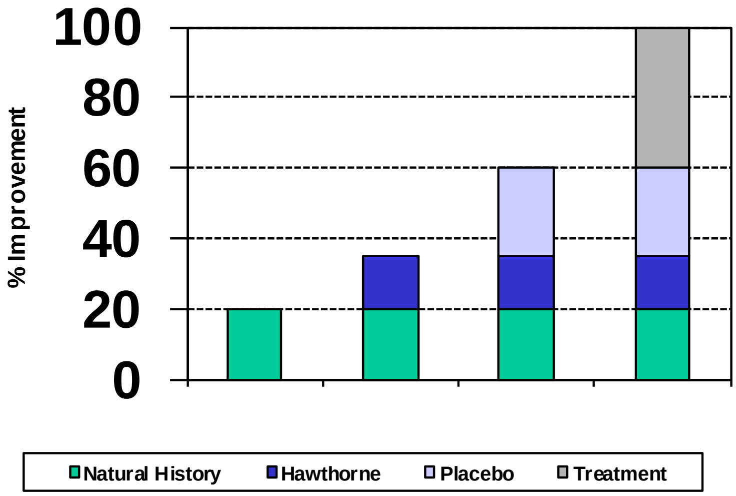
5. Intervention

It is important that the balance between the potential benefits and risks of an intervention be considered carefully before a trial is begun. Remember that everyone is exposed to potential side effects of an intervention, whereas not everyone can benefit from the intervention itself (because not everyone will have or develop the outcome you are trying to prevent and no intervention is ever 100% effective). So caution dictates using the "lowest effective dose". Since RCTs are designed under the premise that serious side effects are expected to occur much less frequently that the outcome, it is not surprising that RCT's are invariably under-powered to detect side effects. This is the reason for Phase IV post-marketing surveillance studies. It is only after the drug has reached the market, and has been used on many, many people, that evidence for rare but serious side effects is found.

[1] Note that the Chapter 8 of the FF text refers to concealment as allocation concealment and includes it under the description of blinding. This is unfortunate in my opinion because concealment has a fundamentally different purpose from other aspects of blinding. Concealment is designed to prevent selection bias, whereas all other forms of blinding are undertaken to reduce measurement bias. So in my opinion these steps should be kept separate. Also note that the two conditions are mutually independent – so it's possible to have a concealed but not-blinded study or an un-concealed but blinded study (further justification for keeping the concepts separate).

All RCT's require a control group. The purpose of the control group is to measure the cumulative effects of all the other factors that can influence the outcome over time – other than the active treatment itself. As shown in the figure below, this includes spontaneous improvements (due to the natural history and the Hawthorne effect), as well as the well documented placebo effect (see below).

Figure 1. Cumulative effects of spontaneous improvements, non-specific responses, and specific treatment effects (from Fletcher)



6. Blinding (a.k.a. masking)

The use of blinding is another cardinal feature of RCT's. Blinding is important because it preserves the benefits of randomization by preventing the *biased assessment of outcomes*. Blinding therefore prevents measurement bias (as opposed to randomization and concealment which prevent confounding bias and selection bias, respectively). Blinding also helps to reduce non-compliance and contamination or cross-overs - especially in the control group (since they are unaware that they are not getting the active treatment). Usually one thinks about either a single-blind study (where either the patient or the physician is blinded) or a double-blind study (where both the patient and the physician are blinded). However, ideally, blinding should occur at all of the following levels:

- Patients
- Caregivers
- Collectors of outcome data (e.g., research assistants or study physicians),
- Adjudicators of the outcome data (e.g., the adjudication committee or study physicians), and
- The data analyst

A placebo is any agent or process that attempts to mask (or blind) the identity of the true active treatment. It is a common feature of drug trials. The value of the placebo, and blinding in general, is especially important when the primary outcome measure being assessed is non-specific or "soft" – for example, patient self-reported outcomes like degree of pain, nausea, or depression. The placebo effect refers to the tendency for such "soft" outcomes to improve when a patient is enrolled in a treatment study, regardless of whether they are actual receiving an active treatment. The placebo effect can be regarded as the baseline against which to measure the effect of the active treatment.

A true placebo may not be justifiable if a known proven treatment is already the standard of care. For example, in a stroke prevention study looking at a new anti-platelet agent, a true placebo control would be very hard to justify given the proven benefit of aspirin (which would be regarded as the minimum standard of care).

Sometimes blinding (and/or the use of a placebo) is just not feasible – for example in a RCT of a surgical intervention it is very difficult to mask who got the surgery and who did not (at least to the patients and caregiver!). In such situations the study is referred to as an "open" trial. Blinding can also be hard to maintain – for example, when the treatment has very clear and obvious benefits, or side effects or other harms. In such cases it is important to try to choose a "hard" outcome – like death from any cause, and to standardize treatments and data collection as much as possible.

7. Loss-to-follow-up, non-compliance, and contamination and missing data

It is important that everyone assigned to either the intervention or control group receives equal follow-up and are ultimately all accounted for. Loss-to-follow-up (LTFU), non-compliance, and contamination are important potential problems in all RCT's. If they occur with any frequency the study will have reduced power (because there are fewer subjects to provide information), and if they occur in a non random fashion (meaning they occur in one arm more than the other - which is often the case) then bias maybe introduced.

Loss to-follow-up can occur for many reasons most of which can be related to the outcomes of interest. For example, subjects are more likely to be lost-to-follow-up if they have side effects, or if they have moved away, got worse, got better or have simply lost interest. Death can be another cause of loss-to-follow-up. If the final outcome of the subjects LTFU remains unknown, then they cannot be included in the final analysis and so (if the rate of loss to-follow-up is high and/or differentially affects one group more than the other) they can have a significant negative effect on the study's conclusions (i.e., low power due to the smaller sample size and biased results due to the differential LTFU). Note that the negative effects of LTFU cannot be easily corrected by the intention-to-treat analysis, since without knowledge of the final outcome status these subjects have to be dropped from the analysis (See below for further details).

Similarly, poor compliance can be expected to be related to the presence of side effects, iatrogenic drug reactions, whether the patient got better or got worse, or whether the patient simply lost interest. People who do not comply with an intervention can be expected to have a worse outcome than those that do. An example of this phenomenon is shown in the table below from the Clofibrate trial – one of the earliest lipid lowering therapy RCTs. The table shows the 5-year cumulative mortality rate by treatment group (as assigned by randomization), and according to whether they complied with the trial protocol. Persons who did not comply had a much higher mortality rate at 5-years, regardless of the treatment group they were assigned to:

Clofibrate Trial	Cumulative 5-year mortality	
	Compliant	Non-compliant
Treatment	15.0%	24.6%
Control	15.1%	28.3%

So the degree to which non-compliance occurs in a study, and the degree to which it differentially affects one arm of the study more than the other, is important to assess.

Contamination refers to the situation when subjects cross-over from one arm of the study into the other – thereby contaminating the initial randomization process. Perhaps the most famous example of this was in the early AIDS treatment RCTs, where participants were able to get their assigned treatments "assayed" by private laboratories to find out whether they were receiving the placebo or active drug (AZT) (the folks in the placebo group would then seek the active AZT drug from other sources!).

As one can imagine excessive loss to follow-up, poor compliance, and/or contamination (see Figure 2) can translate into the slow prolonged death of your trial! Essentially as these three effects take their toll, the original RCT degenerates into a mere observational (cohort) study because the active and compliant participants in each arm of the study are no longer under the control of the study investigator!

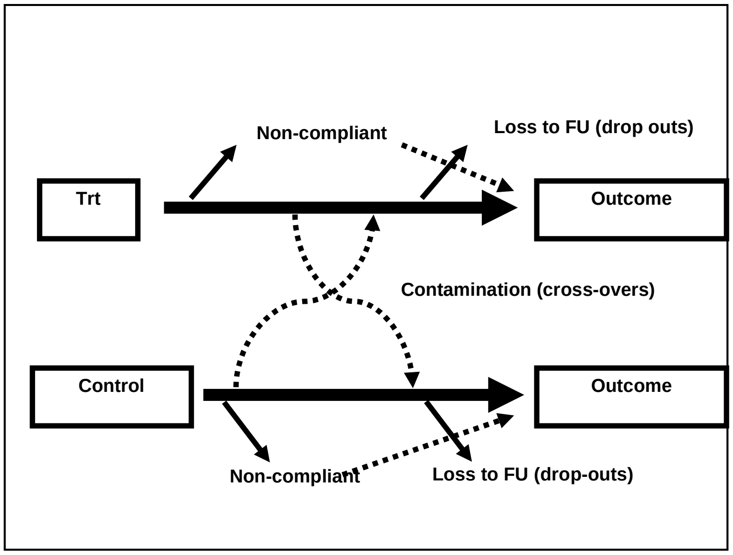
The problem of LTFU is particularly problematic when trials are conducting the gold-standard **intention-to-treat analysis (ITT)** (which is the principle that <u>**all**</u> participants are analyzed according to their original randomization group or arm - regardless of protocol violations). Ideally all subjects should be accounted for in both the denominator and the numerator of the groups event rate. But with lost to follow-up, if the subject is included in the denominator but not the numerator then the event rate in that group will be underestimated. However, if the subjects are simply dropped from the analysis (a common approach) then the analysis can be biased (in fact, if subjects are dropped from the analysis because their outcome status is unknown then one is de facto doing a per protocol (PP) analysis). To mitigate the problems of LTFU some trials will <u>impute</u> an outcome based on a missing data protocol. Techniques for missing data protocols include using the last observation made on the subject (referred to as last observation carried forward), or using the worst observation made on the subject, or using multiple imputation. Multiple imputation uses multivariable statistical models to predict the unobserved outcome based on the specific characteristics of the subjects with missing values; when done properly, it is generally regarded as the best method. But, regardless of the approach used to address missing data, all methods are ultimately unverifiable, and so the results should be viewed with caution. That said some form of imputation is probably better than ignoring the problem of missing data all together unless the amount is small (say, < 5%). Ultimately the best defense is to MINIMIZE missing data through good study design and research practices.

One technique to assess the likely impact of poor compliance or LTFU is the "*5 and 20 rule*" which states that if LTFU or compliance affects < 5% of study participants then bias will be minimal, whereas if it affects >20% then bias is likely to be considerable. One can also assess the potential impact of LTFU by doing a "*best case worst case*" *sensitivity analysis*. In the best case scenario, the subjects LTFU are assumed to have had the best outcome (e.g., none of them had the adverse outcome) and the event rates in each group are calculated after counting all of the LTFU in the denominator but not in the numerator. In the worst case scenario, all of the subjects LTFU are assumed to have had the adverse outcome so the LTFU are counted in both the numerator and the denominator of the event rates. The overall *potential impact* of the LTFU is then gauged by comparing the actual results with the range of findings generated by the sensitivity analysis. Here's an example:

> In a RCT of asthma patients visiting the emergency department (ED), only 100 of 150 subjects assigned to the treatment arm (which involved an educational intervention requiring the subjects to attend 2 sessions at a local asthma treatment center) complied with the treatment and were available for follow-up. The rate of LTFU was therefore 33% clearly exceeding the 5 and 20 rule. The primary outcome, which was defined as a repeat visit to the ED over the next 6 months, occurred in 15% of the 100 subjects who remained in follow-up (i.e., 15 of 100). The results of the best/worst case sensitivity analysis were: best case 15/150 = 10% and worst case 65/150 = 43%. So, clearly the study's findings are questionable given the high loss to follow-up (33%) and the wide range of estimates for the repeat ED visit rate in the treatment arm.

Trialist's go to great lengths to attempt to reduce these problems by enrolling subjects who are more likely to be compliant and not lost-to-follow-up. To do this, studies will sometimes use two screening visits prior to actually performing the enrollment (to weed out the "time-wasters"), and in drug trials will often use pre-randomization run-in periods (using placebo and active drug) to weed out the early non-adheres. Once subjects are enrolled into a study it is very important to minimize LTFU as much as possible. This means that you should attempt to track and follow-up on all the subjects who were non-compliant, ineligible, or chose to drop out of your study for any reason. This is not easy to do!

Figure 2. Effects of non-compliance, loss-to-follow-up, and contamination on a RCT.



8. Measuring Outcomes, Sub-group analyses and Surrogate end points

The primary and secondary study outcomes – with associated definitions should be defined before the study is started (these are termed *a priori* or pre-specified comparisons). The best outcome measures are hard, clinically relevant end points such as disease rates, death, recovery, complications, or hospital/ER use. It is important that all outcomes are measured with accuracy and precision, and, most importantly, that they are measured in the same manner in both groups or arms. It is also important that the outcomes chosen for a trial are clinically relevant to the patients themselves – for example death, recovery, complications - are all important to individual patients. These measures are referred to as **patient-reported outcomes measures (PROMs)**.

Hard patient-relevant clinical outcomes cannot always be used however, for example it usually takes too long to measure disease mortality. To reduce the length and/or the size of the intended study **surrogate end points** may be used under the proviso that they are validated biologically relevant endpoints (often referred to as **biomarkers**). Obviously one should carefully consider whether a surrogate end point used is an adequate measure of the real outcome of interest. Ideally, prior RCTs should be proven that the end point is a valid surrogate measure for the real outcome of interest. For example, in a study designed to reduce stroke incidence, the degree of blood pressure reduction would be considered a valid surrogate end point given the known causal relationship between blood pressure and stroke risk, whereas the reduction in hs-CRP or CRP would not be.

Pre vs. post-hoc sub-group analyses. Sub-group analyses refer to the examination of the primary outcome among study sub-groups defined by key prognostic variables such as age, gender, race, disease severity etc. Sub-group analyses can provide important information by identifying whether the treatment has a different effect within specific sub-populations (differences in the efficacy of a treatment between sub-groups may be described in terms of a ***treatment-by-sub-group interaction***). It is critical that all sub-group analyses be pre-specified ahead of time.

This is because there is a natural tendency among authors of trials that did not show a positive treatment effect for the primary outcome of interest to go "fishing" for positive results by conducting all manner of sub-group comparisons. This approach naturally leads to the statistical problem of *multiple comparisons* and the potential for false positive statistical results (Type 1 errors). Thus all sub-group comparisons which are not pre-specified (i.e., that were *post-hoc*) should be regarded as "exploratory findings" that should be re-examined in future RCT's as pre-panned comparisons.

9. Statistical Analyses

Statistical analyses of trials are on the face of it typically very straight forward, since the design has created balance in all factors, except for the intervention per se. So, oftentimes it's a simple matter of comparing the primary outcome measure between the two arms. For continuous measures the t-test is commonly used, while the Chi-square test is used for categorical outcomes. Non-parametric methods maybe used when the study is relatively small or when the outcomes are not normally distributed. In survival type studies, Kaplan Meire survival curves or advanced Cox regression modeling may be used to study the effect of outcomes which occur over time (and to help assess the problems of loss-to follow-up and censoring).

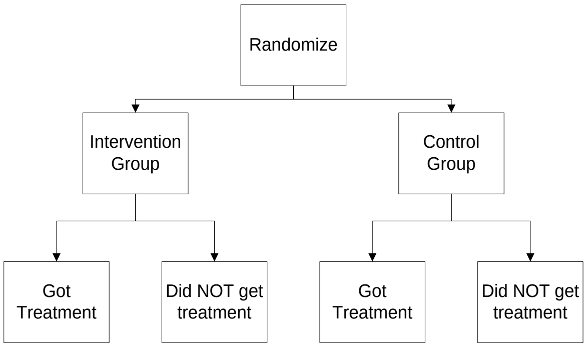
The most important concept to understand in terms of the statistical analysis of RCT's is the principle of **Intention-to-Treat Analysis (ITT)**. This refers to the analysis that compares outcomes based on the *original treatment arm that each individual participant was randomized to regardless of protocol violations.* Protocol violations include ineligibility (i.e., subjects who should not have been enrolled in the study in the first place), non-compliance, contamination or LTFU. The ITT analysis results in the **most valid but conservative estimate** of the true treatment effect. The ITT is the approach that is truest to the principles of randomization - which seeks to create perfectly comparable groups at the outset of the study. It is important to note that not even the ITT analyses can fix the problem of loss-to-follow-up unless the missing outcomes are imputed using a valid method (which can never be fully verified). Thus, in my opinion, no amount of fancy statistics can fix the problem that the final outcome is unknown for a sub-set of subjects.

Two other alternative analytical approaches may be used (as shown the diagram below) but they are both **fundamentally flawed**. The first is the **Per Protocol (PP)** analysis in which subjects in the treatment arm who did not comply with the treatment and control subjects who got treated (i.e., cross-overs) are simply dropped from the analysis. Only those subjects who complied with the original randomization scheme are therefore analyzed. The PP analysis answers the question as to whether the treatment works among those that comply, but it can never provide an unbiased assessment of the true treatment effect (because the decision to comply with treatment is unlikely to occur at random). Also, as mentioned previously, if subjects are dropped from the ITT analysis because their outcome status is unknown, then one is de facto doing a PP analysis.
N.B. alternative names for the PP analysis include *efficacy, exploratory, or effectiveness* analyses.

The second alternative approach is the egregious **As Treated (AT)** analysis, in which subjects are analyzed according to whether they got the treatment or not (regardless of which group they were originally assigned to). So the non-compliant subjects in the intervention arm are moved over to the control arm, and vice versa. This analysis is akin to analyzing the trial as if a cohort study had been done (i.e., everyone had decided for themselves whether to get treated or not), and completely destroys any of the advantages afforded by randomization.

You will see examples of published studies that used the AT analysis (typically they are studies that did not show a positive ITT analysis), but you have to ask– "what was the point of doing the trial in the first place if you ended up doing an AT analysis?" The AT approach is without merit![2]

The analysis of RCT's - Intention-To-Treat versus Per Protocol or As Treated



Interntion-to-Treat	YES	YES	NO	NO
Per protocol	YES	DROP	DROP	NO
As Treated	YES	NO	YES	NO

<u>Where</u>:
YES means that the group is included in the analysis as the group that got treatment.
NO means that the group is included in the analysis as the group that DID NOT get treatment.
DROP means that the group is not included in the analysis (it is simply ignored).

In the table below is a real life example of the application of these three analytical approaches to a trial that compared surgical treatment (CABG) to medical treatment for stable angina pectoris in 768 men (European Coronary Surgery Study Group. Lancet 1979;i:889-93). 373 men were randomized to medical treatment but 50 ended up being treated by surgery. Of the 394 subjects randomized to surgery treatment, 26 did not receive it. A further subject was lost to follow-up and was dropped from these analyses which are based on 767 subjects. The death rates according to the actual treatments received are calculated and then the absolute risk reduction (ARR) (for surgery vs. medical) with 95% CI are calculated using the three different approaches.

Note the very high death rate in the 26 subjects who should have gotten surgery but received medical treatment – clearly these are a very sick group of patients who either died before surgery or were too sick to undergo it. Note also the 52 subjects who should have gotten medical treatment but somehow got surgery. Their mortality (4%) is much lower than the rest of the medically treated group (8.4%) – however, these men could have been healthier at baseline and so its impossible to judge the relative merits of surgery based on these two figures.

[2] The FF text refers to this analysis as an explanatory trial, which in my opinion this does not disparage this approach sufficiently!

Table: Effect of different analysis approaches on an RCT of coronary artery bypass surgery versus medical treatment in 767 men with stable angina. (Lancet 1979;i:889-93).

	Allocated (vs. actual) treatment				
	Medical (medical)	Medical (surgical)	Surgical (surgical)	Surgical (medical)	ARR (95% CI)
Num. subjects	323	50	368	26	
Deaths	27	2	15	6	
Mortality (%)	8.4%	4.0%	4.1%	23.1%	
ITT analysis	7.8% (29/373)		5.3% (21/394)		2.4% (-1.0, 6.1)
PP analysis	8.4% (27/323)		4.1% (15/368)		4.3% (0.7, 8.2)
AT analysis	9.5% (33/349)		4.1% (17/418)		5.4% (1.9, 9.3)

When subjects were analyzed according to the groups they were randomized to (ITT), the results show that surgery had a small non-significant benefit vs. medical treatment. However, when the data was analyzed according to those that complied (PP), or according to the final treatment received (AT), the results show a larger and now statistically significant benefit for surgery. However, both estimates are **biased** in favour of surgery because the 26 high risk subjects were either dropped from the surgery group or were moved into the medical group – so clearly this analysis is stacked against medical treatment!!

10. RCTs and Meta-analyses - assessing trial quality, trial reporting, and trial registration

Meta-analyses of RCTs have fast become the undisputed king of the evidence-based tree. The preeminence of meta-analysis as a technique has had three important implications for the RCT.

i) Assessment of study quality – Given the variability in the quality of published RCTs, meta-analysts will often attempt to assess their quality in order to determine whether the quality of a trial has an impact on the overall results. While there are several approaches to conducting quality assessment of RCTs (example: Jadad, 1996), they all essentially focus on the same criteria: a description of the randomization process, the use of concealment, the use of blinding, and a description of the loss-to-follow-up and non-compliance rates.

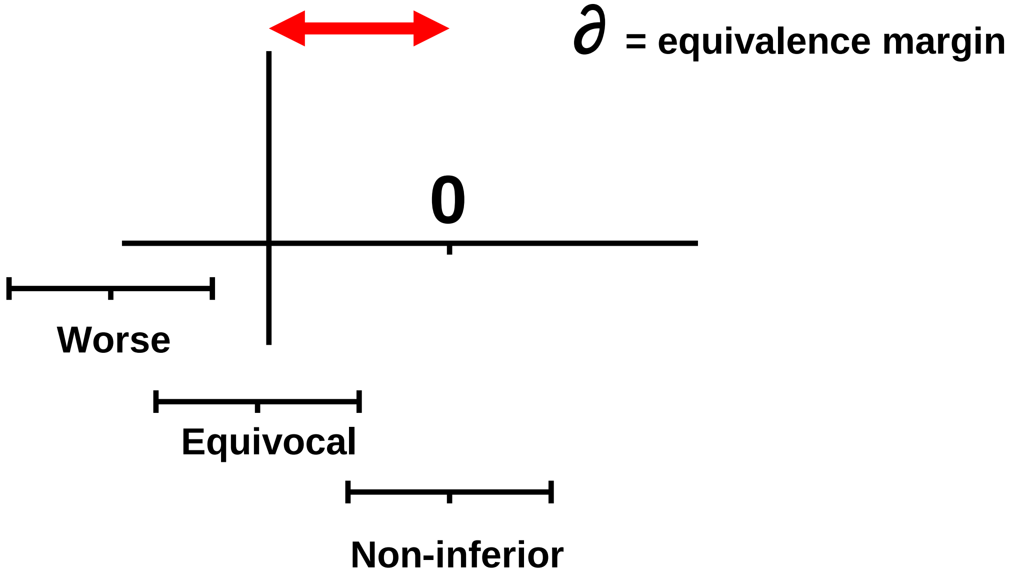
ii) Trial reporting – Quality assessment of published trials using the Jadad scale or other similar tools often indicates that the trials are of marginal or poor quality – in part, because they did not report information on the key quality criteria (i.e., randomization, concealment, blinding, LTFU). One is therefore left unsure as to whether the trial did not follow these basic steps, or whether the authors simply failed to report these steps in the paper. This problem has led to the development of specific guidelines for the reporting of clinical trials, called the CONSORT Statement (see Moher, JAMA 2001). This statement aims to make sure that trials are reported in a consistent fashion and that specific descriptions are included so the validity criteria can be independently assessed.

iii) Trial registration – A big problem for meta-analyses is the potential for publication bias. The results of any meta-analysis can be seriously biased if there is a tendency to not publish trials with negative or null results. Unpublished negative trials have the tendency to be either small studies (that had insufficient power to detect a clinically important difference), or on occasion, large trials that were sponsored by drug companies who do not want to release the information that shows their drug or gadget did not live up to their expectations. Several controversies in the past few years has led to the International Committee of Medical Journal Editors (which includes JAMA, NEJM, Lancet etc) to require that for a trail to be published in any of these journals it must have been registered prior to starting it!. The idea here is that the scientific community will then have a registry of all trials undertaken on a topic, rather than the current situation where the published literature represents all trials that were deemed worthy of publication.

Equivalence and Non-inferiority Designs

For many conditions there exists a standard treatment which makes the use of a placebo-controlled trial not ethically acceptable. Therefore new drugs need to be compared to this active control. But it is increasingly difficult to prove that a new drug is **better** than an existing drug. Thus, an alternative approach is to prove that new drug is **no worse** than active control (within a given tolerance (∂) or equivalence margin. There is now increasing emphasis at the federal level on the conduct of *comparative effectiveness* trials i.e., trials done to directly compare alternative treatments. Comparative effectiveness trials use either equivalence or non-inferiority designs.

Non-inferiority Trials



<u>Equivalence trials</u> are trials designed to prove that a new drug is equivalent to an existing standard drug within a given tolerance (∂) or equivalence margin. They are most often used in the area of generic drug development to prove that a new generic drug is bio-equivalent to the original drug i.e., similar bioavailability, pharmacology etc.

Non-inferiority trials are trials designed to prove that a new drug is no less effective than an existing standard drug. This is a one-sided equivalence test. The interest in non-inferiority trials assumes that the new drug has other advantages:

Better safety profile - less side effects, less monitoring
Easier dosing schedule - better compliance
Cheaper

Non-inferiority trials may therefore involve the evaluation of the same drug given using a different strategy, dose or duration.

There are several methodological challenged associated with non-inferiority trials. First is that the null hypothesis being testing is opposite to that of a typical superiority trial. In a superiority trial the null and alternative hypotheses are:

H_0: New drug = Active Control vs. H_A: New drug $\neq$ Active Control.

Whereas in the non-inferiority trial the null and alternative hypotheses are:

H_0: New drug + ∂ < Active Control vs. H_A: New drug + $\partial \geq$ Active Control (where ∂ is the equivalence margin).

The null hypothesis for the non-inferiority trial is that the active control is substantially better than the new drug. Accepting the null hypothesis means that the new drug is worse that active control by more than the equivalence margin (**∂)**. Rejecting the null hypothesis means that the new drug is not inferior to the active control within the bounds of the equivalence margin.

The equivalence margin (∂) (which is also referred to as tolerance or the non-inferiority margin) indicates by how much we are willing to accept that the new drug can have worse efficacy. The margin is set by deciding clinically on how big a difference there would have to be between the two drugs before we would decide that the drugs are clinically not equivalent i.e., that you would prefer one over the other. ∂ is the critical determinant of the success of the trial and its sample size. Smaller values of ∂ are more conservative, larger values more liberal.

Other problems and limitations of non-inferiority trials

Assay sensitivity - A poorly conducted trial may falsely show that the 2 drugs are equivalent. Poor trial conduct in terms of compliance, follow-up, blinded assessments etc will favour non-inferiority.

Blinding – in superiority trials this is a vital step to reduce measurement bias. However, blinding does not have the same potential for benefit in a non-inferiority trial since it cannot protect against the investigators giving the same outcomes/ratings to all subjects and thereby showing non-inferiority.

ITT analysis - In superiority trials the ITT is regarded as the gold standard analysis approach. But doing an ITT analysis for non-inferiority trials tends to bias towards finding non-inferiority! Including the non-compliant subjects in the treatment and active control groups tends to minimize differences between the groups thus potentially showing an inferior drug to be non-inferior. Compounding this problem is the fact that doing a PP analysis can introduce bias in either direction, hence it is not recommended. The best approach is to do both an ITT and PP analysis and hope that the findings are consistent. But even then accepting the Ha of non-inferiority does not rule out the possibility of bias.

III. Advantages and disadvantages of RCT's - Summary

Advantages:

High internal validity

Control of exposure – including the amount, timing, frequency, duration

Randomization - ensures balance of factors that could influence outcome i.e., it "controls" the effect of known and unknown confounders

A true measure of efficacy.

Disadvantages:

Limited external validity

Artificial environment, that includes the strict eligibility criteria and the fact that they are conducted in specialized tertiary care (referral) medical centers limits generalizability (external validity).

Difficult/complex to conduct, take time, and are expensive.

Because of ethical considerations they have limited scope – mostly therapeutic / preventive only

Finally, one should be sure to re-familiarize yourself with the measures of effect that are commonly used to describe the results of RCT's. These were covered in the Frequency lecture and include the CER or baseline risk, the TER, the RR, the RRR, the ARR and the NNT.

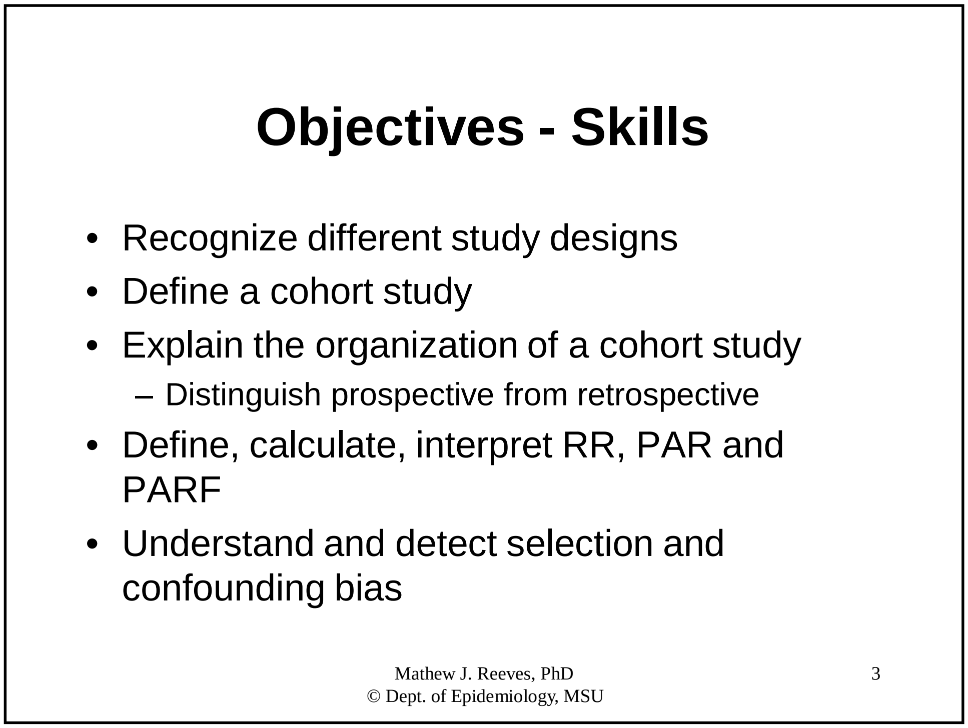
EPI-546 Block I

Lecture – Study Design I XS and Cohort Studies

Mathew J. Reeves BVSc, PhD
Associate Professor, Epidemiology

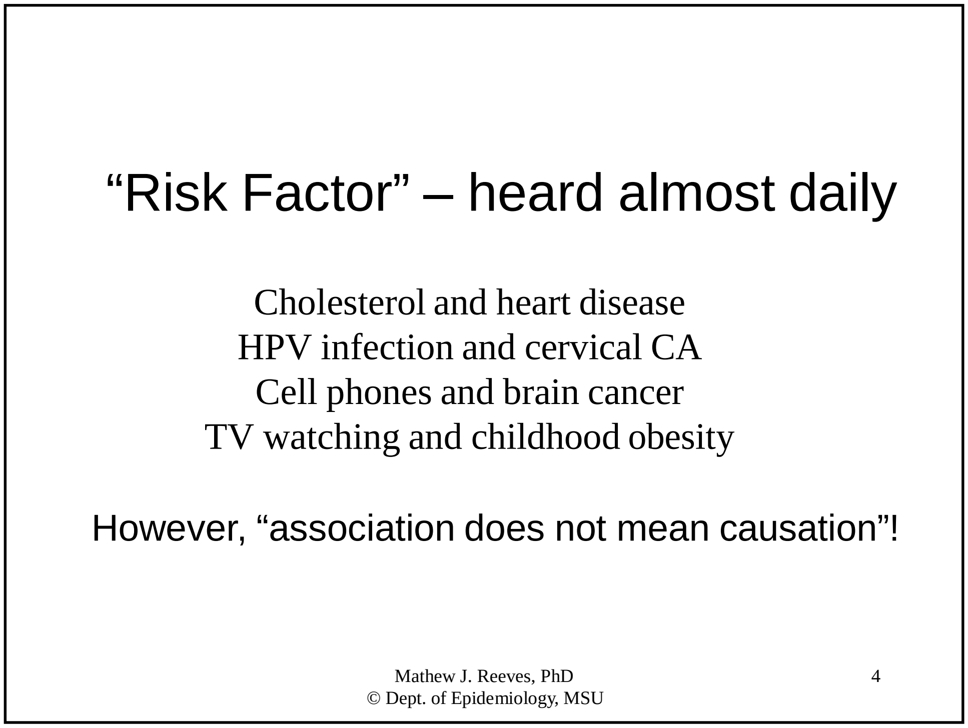
Objectives - Concepts

- Uses of risk factor information
- Association vs. causation
- Architecture of study designs (Grimes I)
- Cross sectional (XS) studies
- Cohort studies (Grimes II)
- Measures of association – RR, PAR, PARF
- Selection and confounding bias
- Advantages and disadvantages of cohort studies



Objectives - Skills

- Recognize different study designs
- Define a cohort study
- Explain the organization of a cohort study
  - Distinguish prospective from retrospective
- Define, calculate, interpret RR, PAR and PARF
- Understand and detect selection and confounding bias



“Risk Factor” – heard almost daily

Cholesterol and heart disease
HPV infection and cervical CA
Cell phones and brain cancer
TV watching and childhood obesity

However, “association does not mean causation”!

Why care about risk factors?

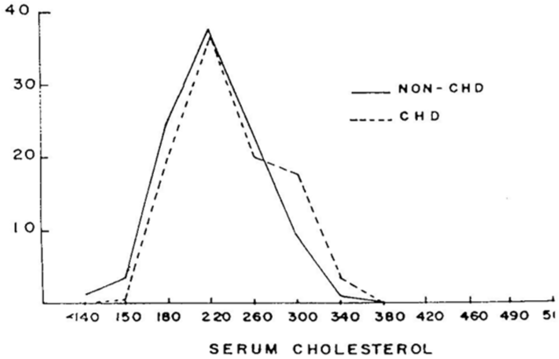
Fletcher lists the ways risk factors can be used:

- Identifying individuals/groups “at-risk”
  - But ability to predict future disease in individual patients is very limited even for well established risk factors
  - e.g., cholesterol and CHD
- Causation (causative agent vs. marker)
- Establish pretest probability (Bayes’ theorem)
- Risk stratification to identify target population
  - Example: Age > 40 for mammography screening
- Prevention
  - Remove causative agent & prevent disease

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU
5

Predicting disease in individual patients

Fig. Percentage distribution of serum cholesterol levels (mg/dl) in men aged 50-62 who did or did not subsequently develop coronary heart disease (Framingham Study)



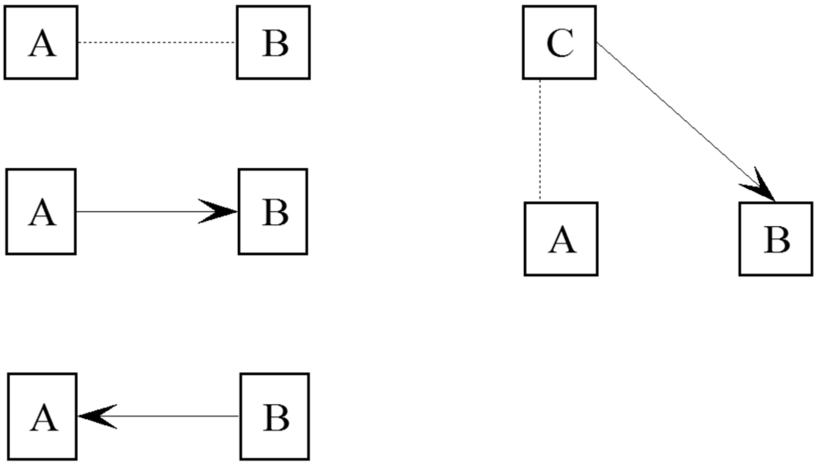
Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU
6

Causation vs. Association

- An association between a risk factor and disease can be due to:
  - the risk factor being a cause of the disease (= a *causative agent*) OR
  - the risk factor is NOT a cause but is merely associated with the disease (= a *marker*)
- Must guard against thinking that A causes B when really B causes A (*reverse causation*).
  - e.g. sedentariness and obesity.

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

7



Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

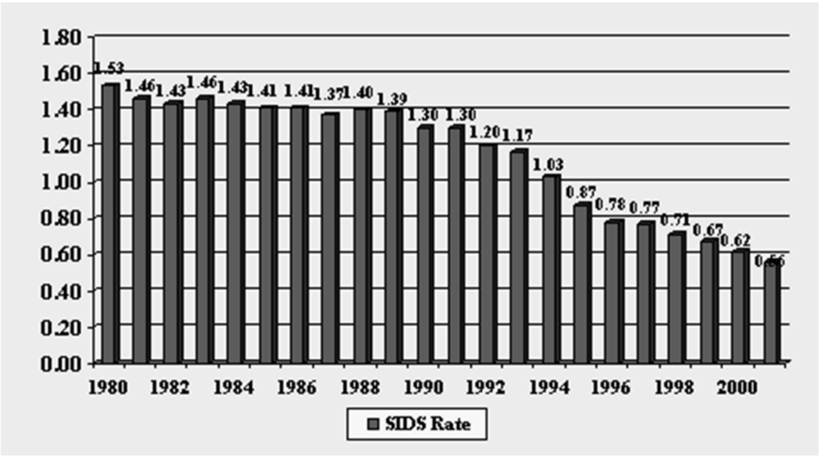
8

Prevention

- Removing a true cause → ↓ disease incidence.
  - – Decrease aspirin use → ↓ Reye’s Syndrome
  - – Discourage prone position

    “Back to Sleep” → ↓ SIDS

Back-To-Sleep Campaign Began in 1992



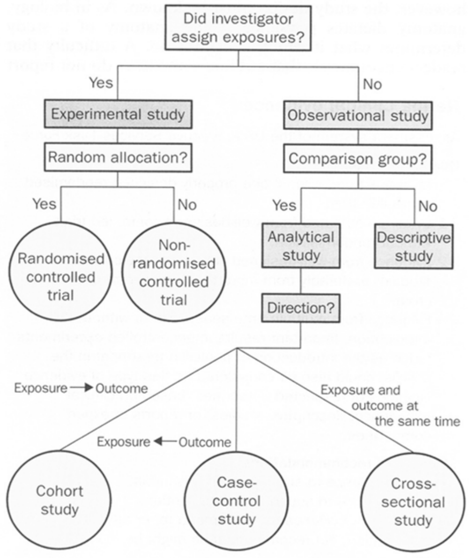
Architecture of study designs

- Experimental vs. observational
- Experimental studies
  - Randomization?
    - RCT vs. quasi-randomized or natural experiments
- Observational studies
  - Analytical vs. descriptive
  - Analytical
    - XS, Cohort, CCS
  - Descriptive
    - Case report, case series

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

11

Grimes DA and Schulz KF 2002. An overview of clinical research. Lancet 359:57-61.



12

Figure 1: **Algorithm for classification of types of clinical research**

Grimes DA and Schulz KF 2002. An overview of clinical research. Lancet 359:57-61.

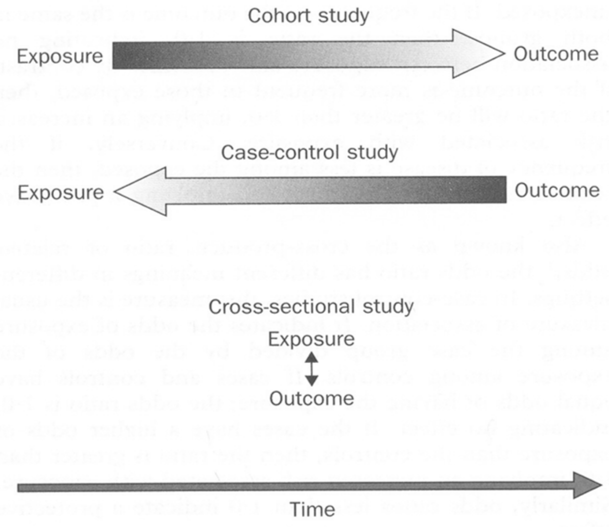


Figure 2: **Schematic diagram showing temporal direction of three study designs**

13

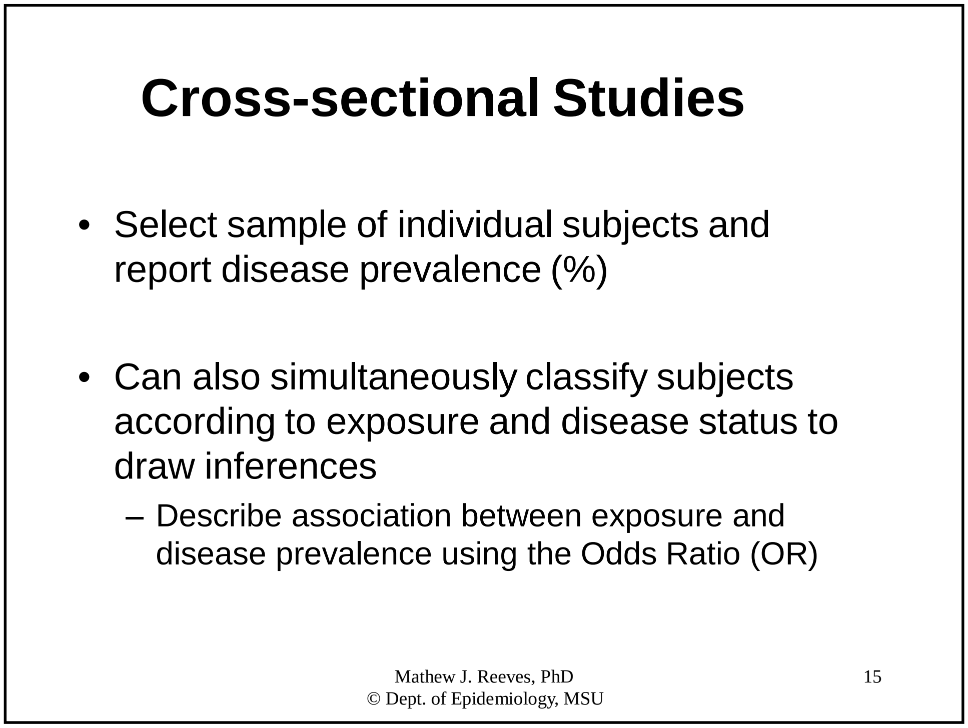
© Dept. of Epidemiology, MSU

Cross-sectional studies

- Also called a *prevalence study*
- Prevalence measured by conducting a <u>survey</u> of the population of interest e.g.,
  - Interview of clinic patients
  - Random-digit-dialing telephone survey
- Mainstay of *descriptive epidemiology*
  - patterns of occurrence by time, place and person
  - estimate disease frequency (prevalence) and time trends
- Useful for:
  - program planning
  - resource allocation
  - generate hypotheses

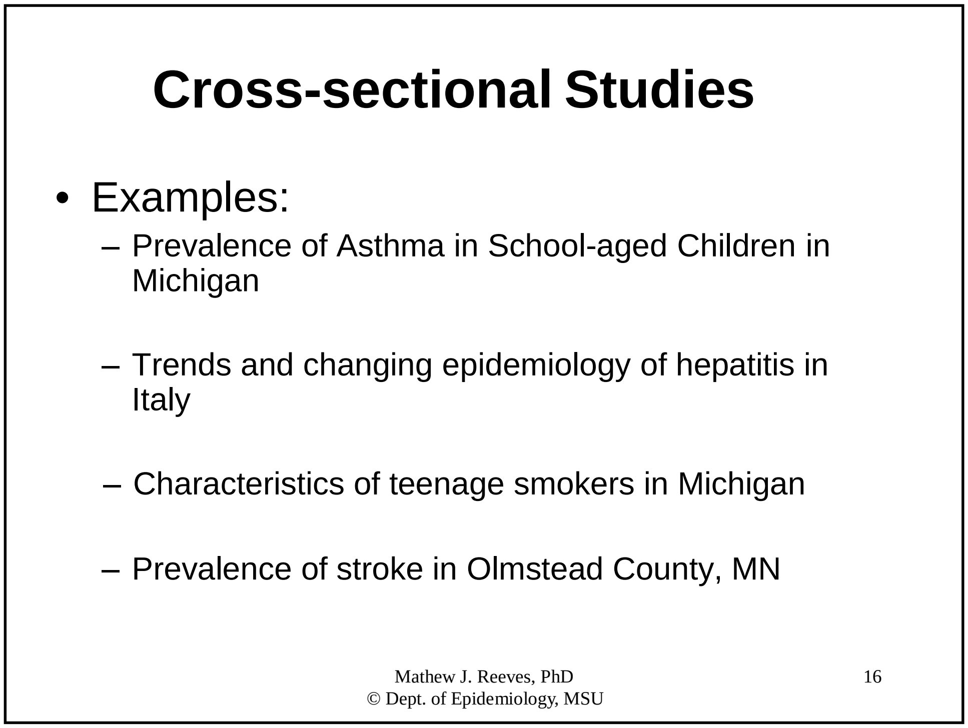
Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

14



Cross-sectional Studies

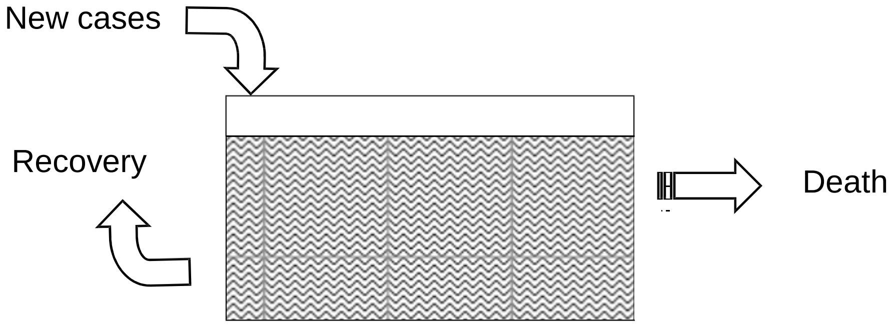
- Select sample of individual subjects and report disease prevalence (%)
- Can also simultaneously classify subjects according to exposure and disease status to draw inferences
  - Describe association between exposure and disease prevalence using the Odds Ratio (OR)



Cross-sectional Studies

- Examples:
  - Prevalence of Asthma in School-aged Children in Michigan
  - Trends and changing epidemiology of hepatitis in Italy
  - Characteristics of teenage smokers in Michigan
  - Prevalence of stroke in Olmstead County, MN

Concept of the Prevalence "Pool"



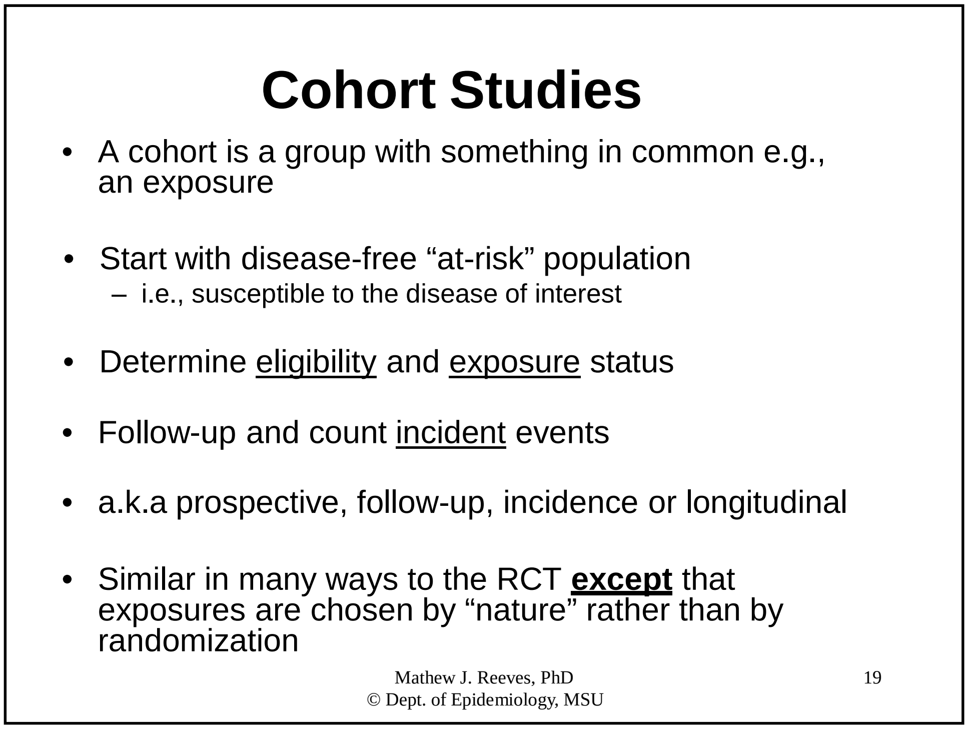
New cases

Recovery

Death

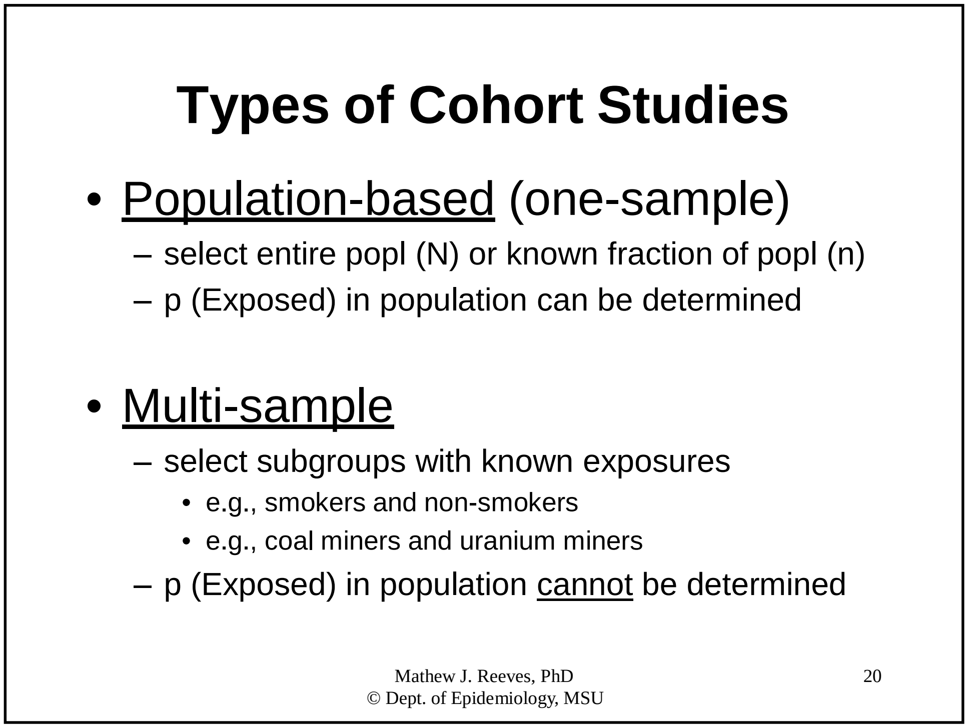
Cross-sectional Studies

- Advantages:
  - – quick, inexpensive, useful
- Disadvantages:
  - – uncertain temporal relationships
  - – survivor effect
  - – low prevalence due to
    - rare disease
    - short duration



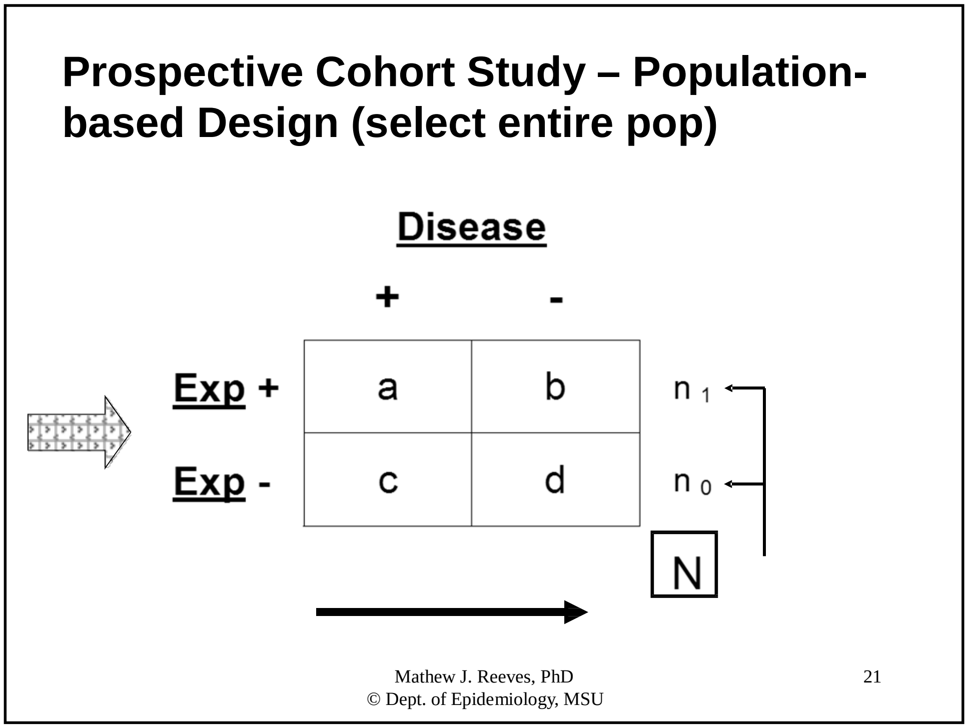
Cohort Studies

- A cohort is a group with something in common e.g., an exposure
- Start with disease-free “at-risk” population
  - i.e., susceptible to the disease of interest
- Determine <u>eligibility</u> and <u>exposure</u> status
- Follow-up and count <u>incident</u> events
- a.k.a prospective, follow-up, incidence or longitudinal
- Similar in many ways to the RCT **<u>except</u>** that exposures are chosen by “nature” rather than by randomization



Types of Cohort Studies

- <u>Population-based</u> (one-sample)
  - select entire popl (N) or known fraction of popl (n)
  - p (Exposed) in population can be determined
- <u>Multi-sample</u>
  - select subgroups with known exposures
    - e.g., smokers and non-smokers
    - e.g., coal miners and uranium miners
  - p (Exposed) in population <u>cannot</u> be determined



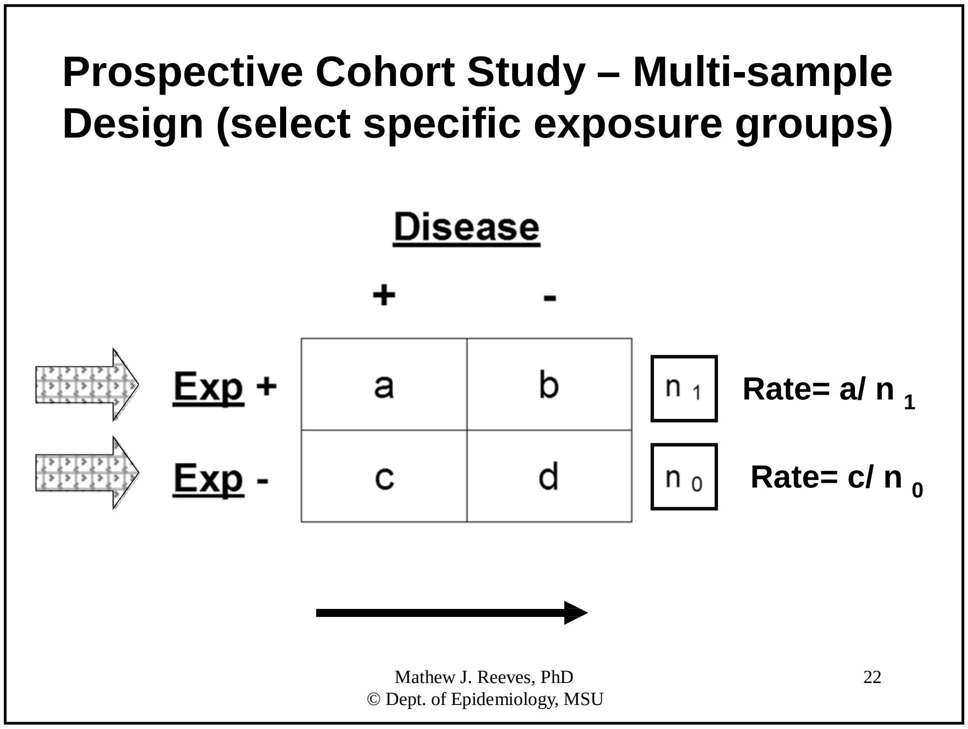
Prospective Cohort Study – Population-based Design (select entire pop)

Disease

	+	-	
Exp +	a	b	n_1
Exp -	c	d	n_0
			N

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

21



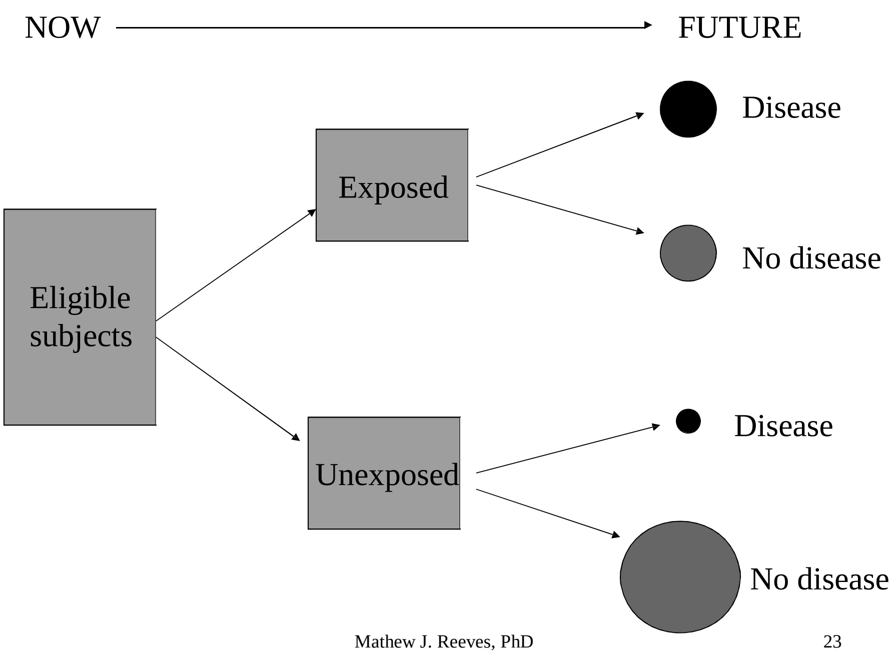
Prospective Cohort Study – Multi-sample Design (select specific exposure groups)

Disease

	+	-		
Exp +	a	b	n_1	Rate= a/ n_1
Exp -	c	d	n_0	Rate= c/ n_0

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

22



NOW → FUTURE

Eligible subjects

Exposed

Unexposed

Disease

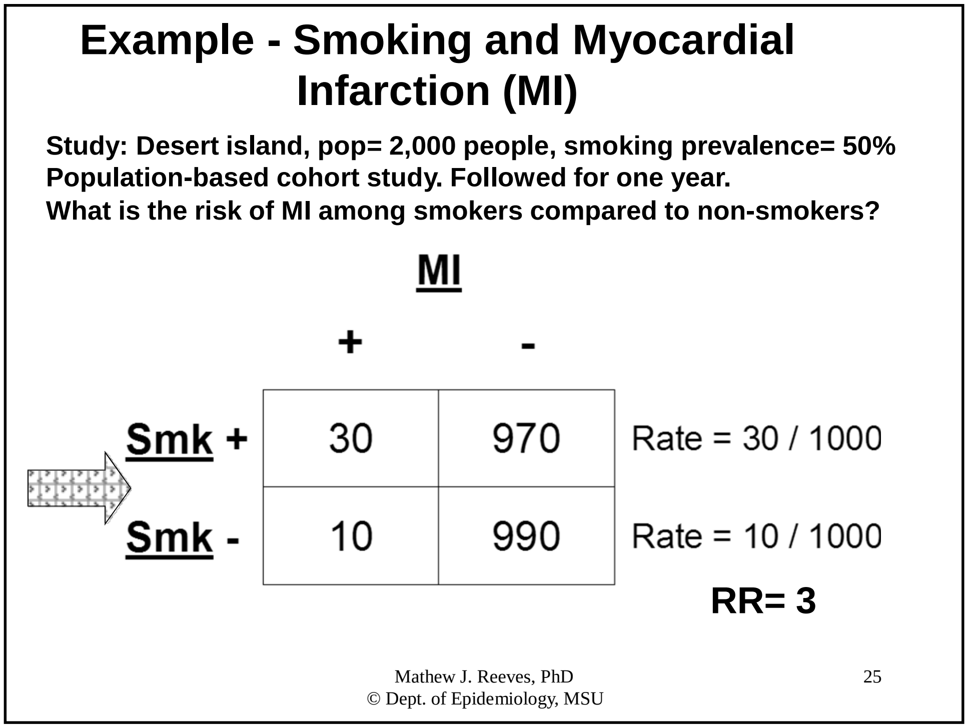
No disease

Disease

No disease

Relative Risk – Cohort Study

- RR = $\dfrac{\text{Incidence rate in exposed}}{\text{Incidence rate in non-exposed}}$
- The RR is the standard measure of association for cohort studies
- RR describes magnitude and direction of the association
- Incidence can be measured as the IDR or CIR
- RR = $\dfrac{a / n_i}{c / n_o}$ or $\dfrac{a / (a + b)}{c / (c + d)}$



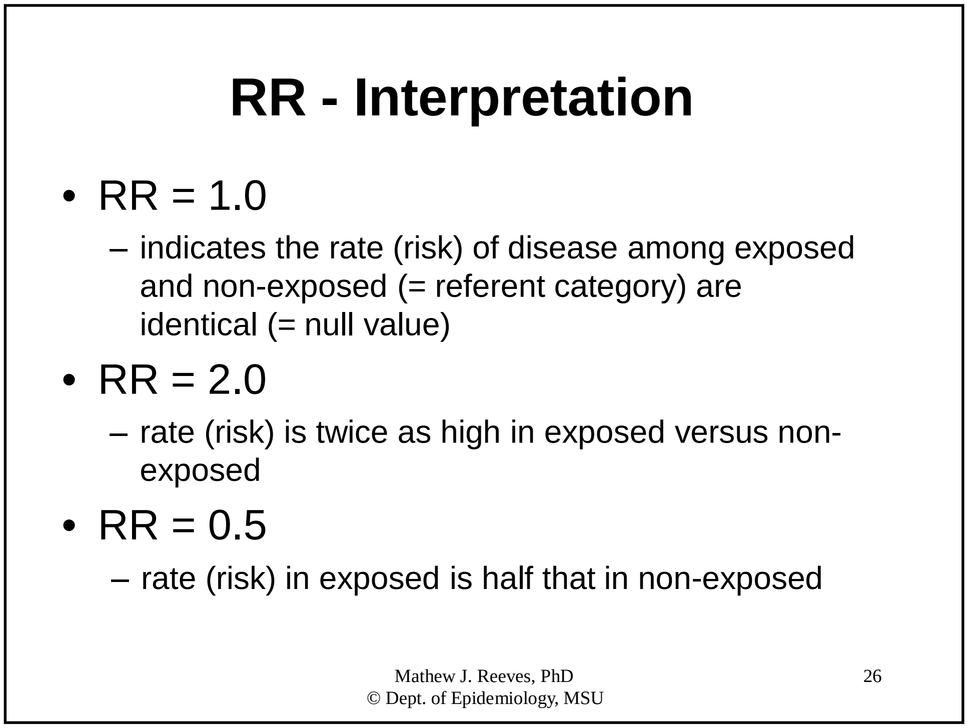
Example - Smoking and Myocardial Infarction (MI)

Study: Desert island, pop= 2,000 people, smoking prevalence= 50%
Population-based cohort study. Followed for one year.
What is the risk of MI among smokers compared to non-smokers?

MI

	+	-	
Smk +	30	970	Rate = 30 / 1000
Smk -	10	990	Rate = 10 / 1000

RR= 3



RR - Interpretation

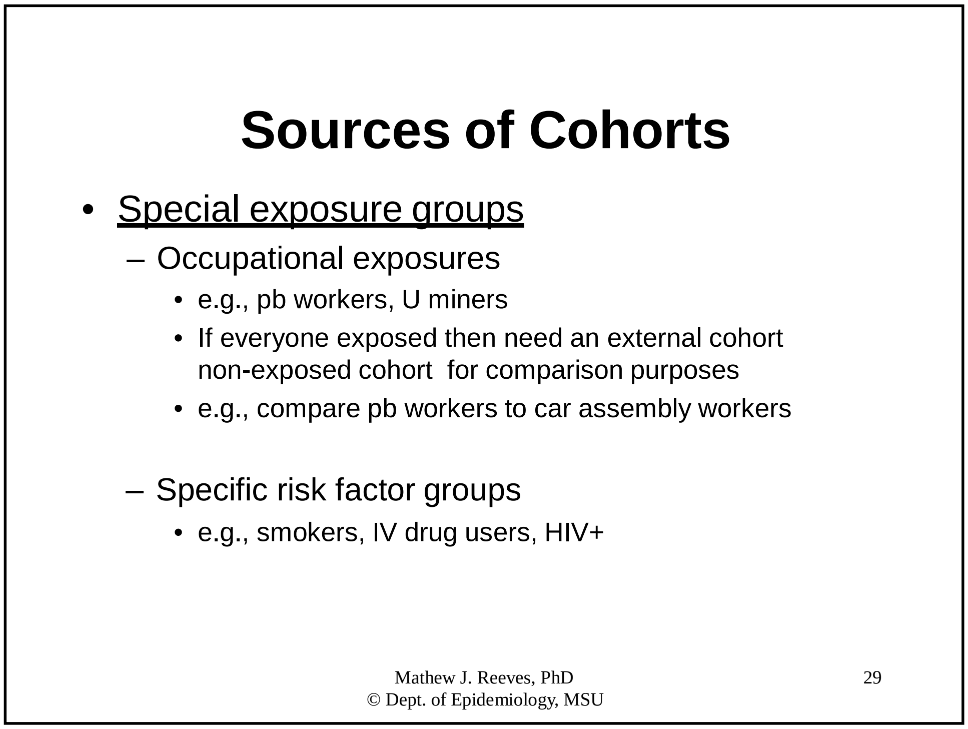
- RR = 1.0
  - indicates the rate (risk) of disease among exposed and non-exposed (= referent category) are identical (= null value)
- RR = 2.0
  - rate (risk) is twice as high in exposed versus non-exposed
- RR = 0.5
  - rate (risk) in exposed is half that in non-exposed

RR – Interpretation (Cohort Studies)

- RR = > 5.0 or < 0.2
  - BIG
- RR = 2.0 – 5.0 or 0.5 – 0.2
  - MODERATE
- RR = <2.0 or >0.5
  - SMALL

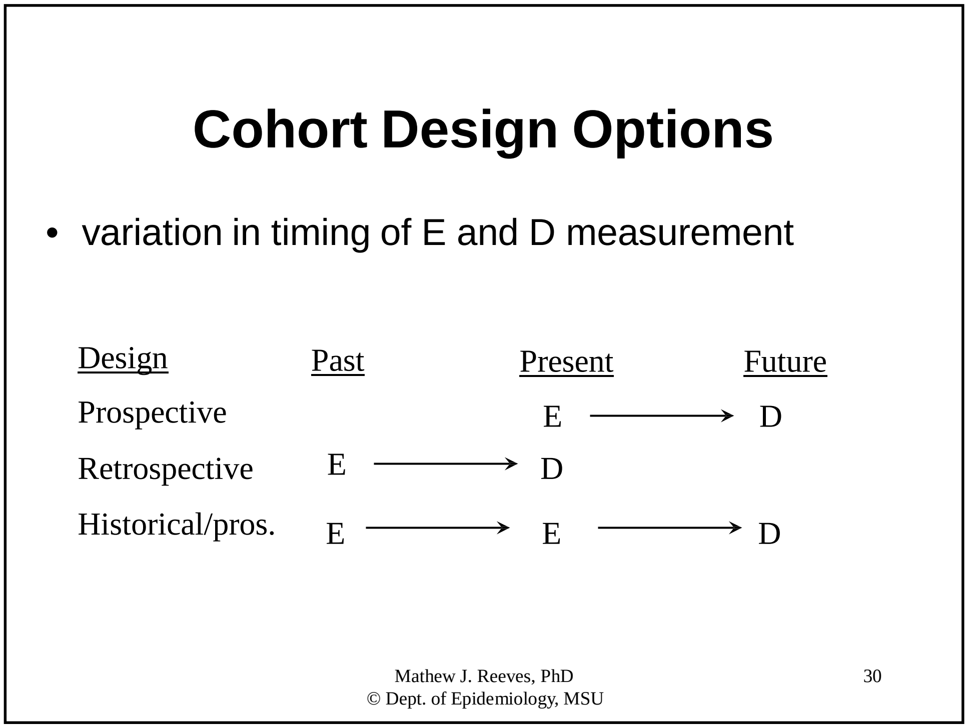
Sources of Cohorts

- <u>Geographically defined groups:</u>
  - Framingham, MA (sampled 6,500 of 28,000, 30-50 yrs of age)
  - Tecumseh, MI (8,641 persons, 88% of population)
- <u>Special resource groups</u>
  - Medical plans e.g., Kaiser Permanente
  - Medical professionals e.g.,
    - Physicians Health Study, Nurses Health Study
  - Veterans
  - College graduates e.g., Harvard Alumni



Sources of Cohorts

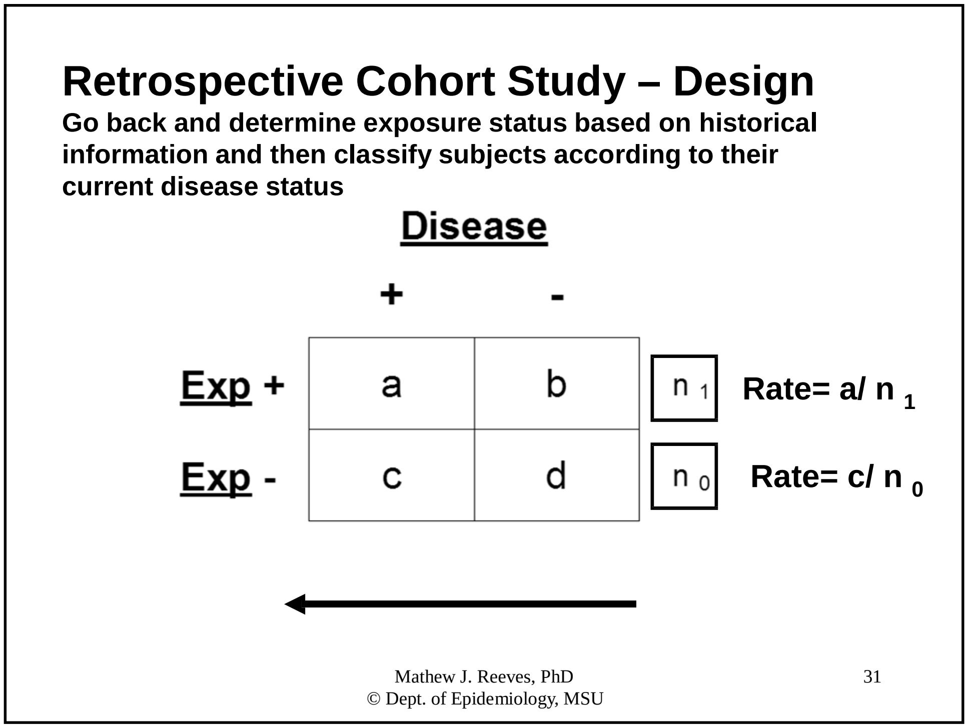
- Special exposure groups
  - Occupational exposures
    - e.g., pb workers, U miners
    - If everyone exposed then need an external cohort non-exposed cohort for comparison purposes
    - e.g., compare pb workers to car assembly workers
  - Specific risk factor groups
    - e.g., smokers, IV drug users, HIV+



Cohort Design Options

- variation in timing of E and D measurement

Design	Past	Present	Future
Prospective		E ⟶	D
Retrospective	E ⟶	D	
Historical/pros.	E ⟶	E ⟶	D



Retrospective Cohort Study – Design

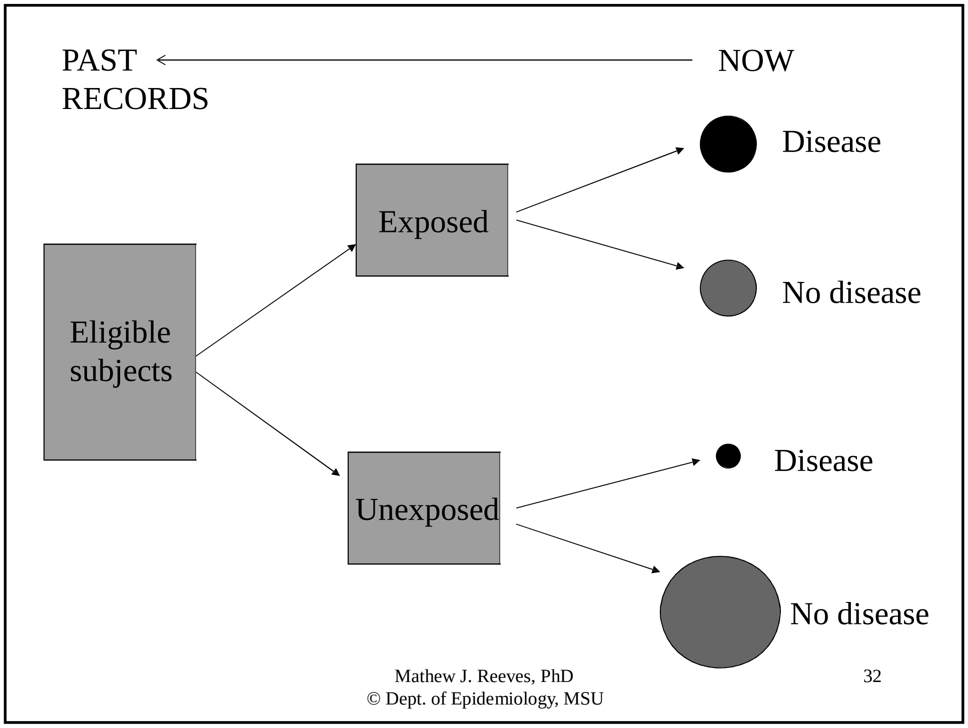
Go back and determine exposure status based on historical information and then classify subjects according to their current disease status

Disease

	+	-		
Exp +	a	b	n_1	**Rate= a/ n_1**
Exp -	c	d	n_0	**Rate= c/ n_0**

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

31



PAST ← NOW
RECORDS

Eligible subjects

Exposed

Unexposed

Disease

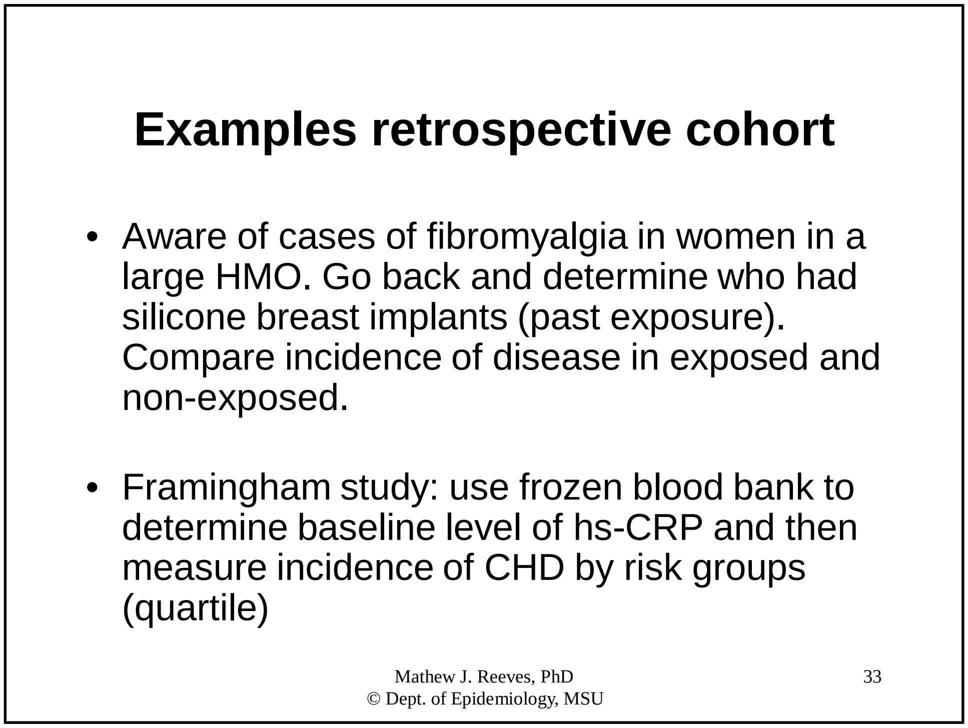
No disease

Disease

No disease

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

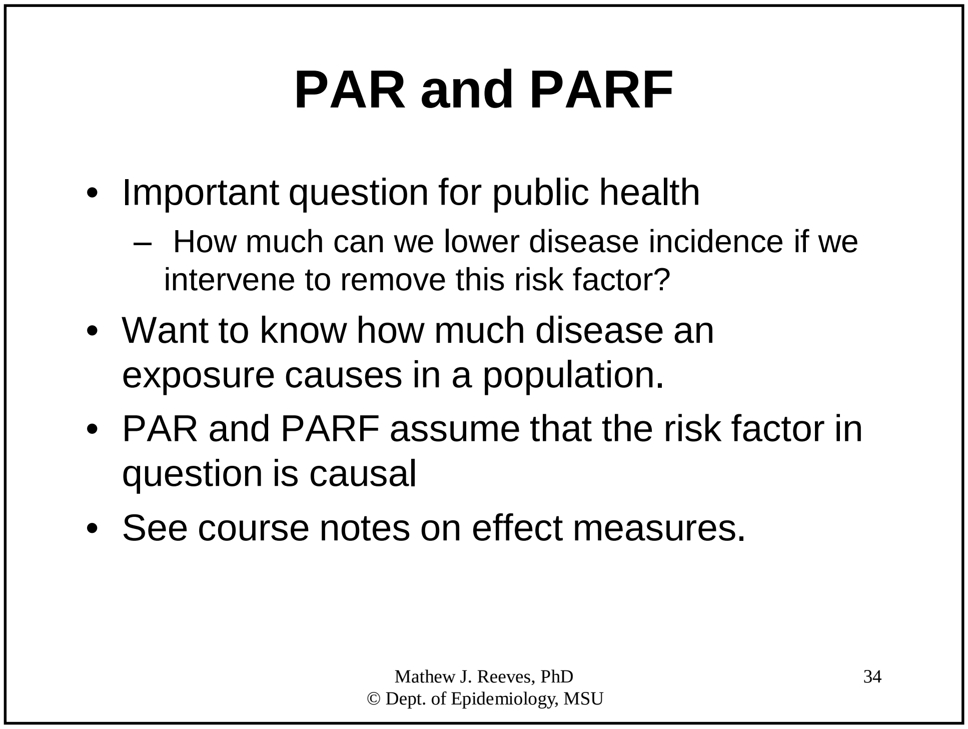
32



Examples retrospective cohort

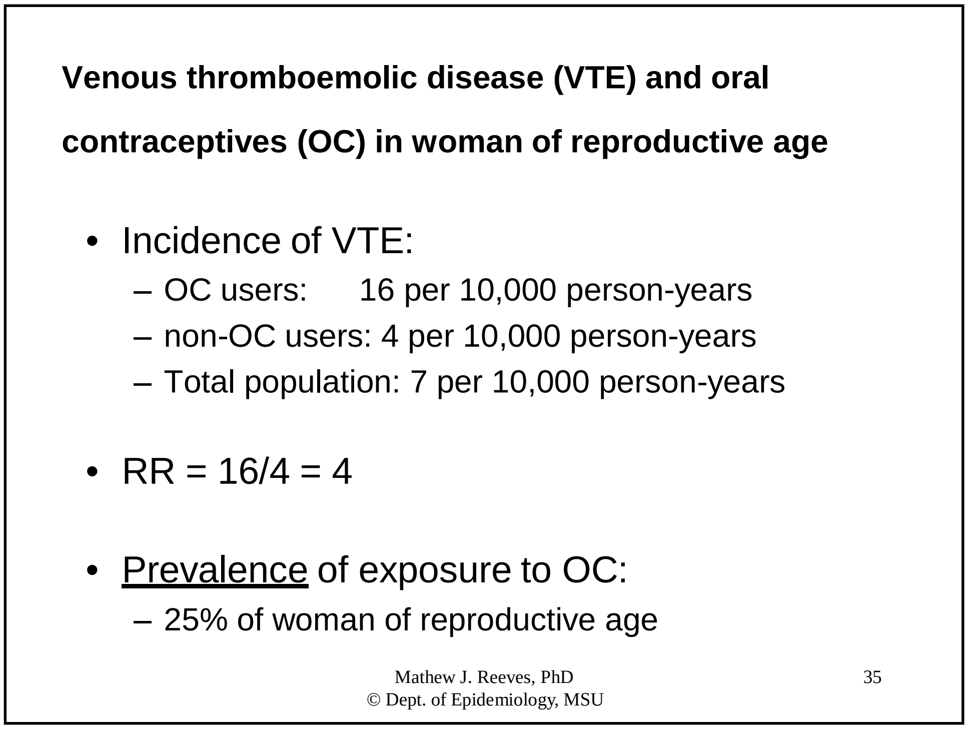
- Aware of cases of fibromyalgia in women in a large HMO. Go back and determine who had silicone breast implants (past exposure). Compare incidence of disease in exposed and non-exposed.

- Framingham study: use frozen blood bank to determine baseline level of hs-CRP and then measure incidence of CHD by risk groups (quartile)



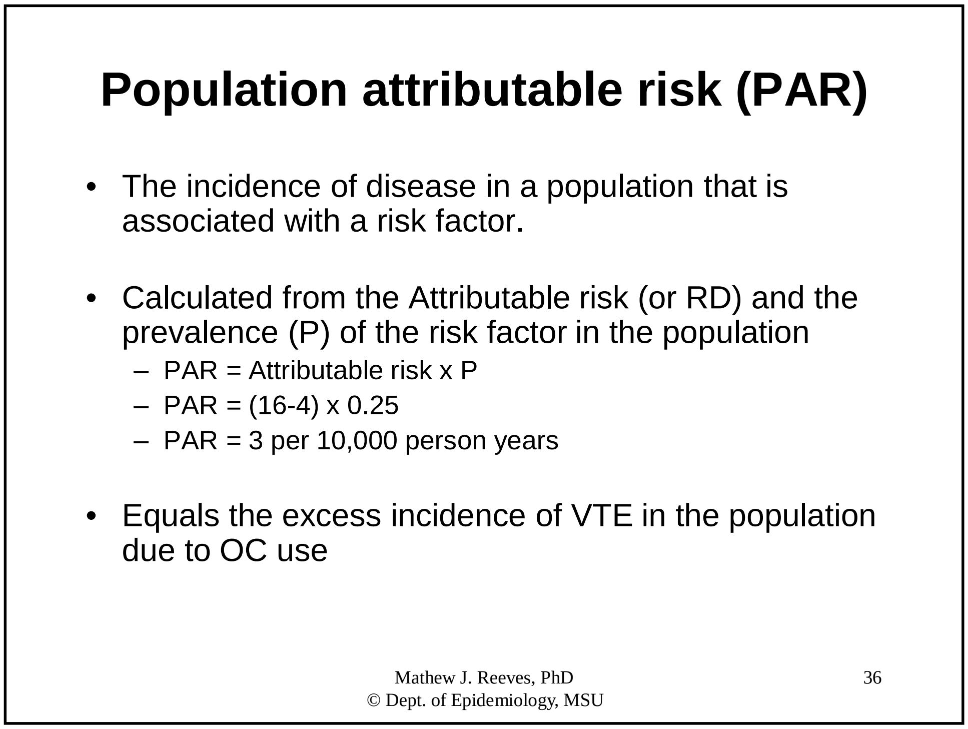
PAR and PARF

- Important question for public health
  - How much can we lower disease incidence if we intervene to remove this risk factor?
- Want to know how much disease an exposure causes in a population.
- PAR and PARF assume that the risk factor in question is causal
- See course notes on effect measures.



Venous thromboemolic disease (VTE) and oral contraceptives (OC) in woman of reproductive age

- Incidence of VTE:
  - OC users: 16 per 10,000 person-years
  - non-OC users: 4 per 10,000 person-years
  - Total population: 7 per 10,000 person-years
- RR = 16/4 = 4
- <u>Prevalence</u> of exposure to OC:
  - 25% of woman of reproductive age



Population attributable risk (PAR)

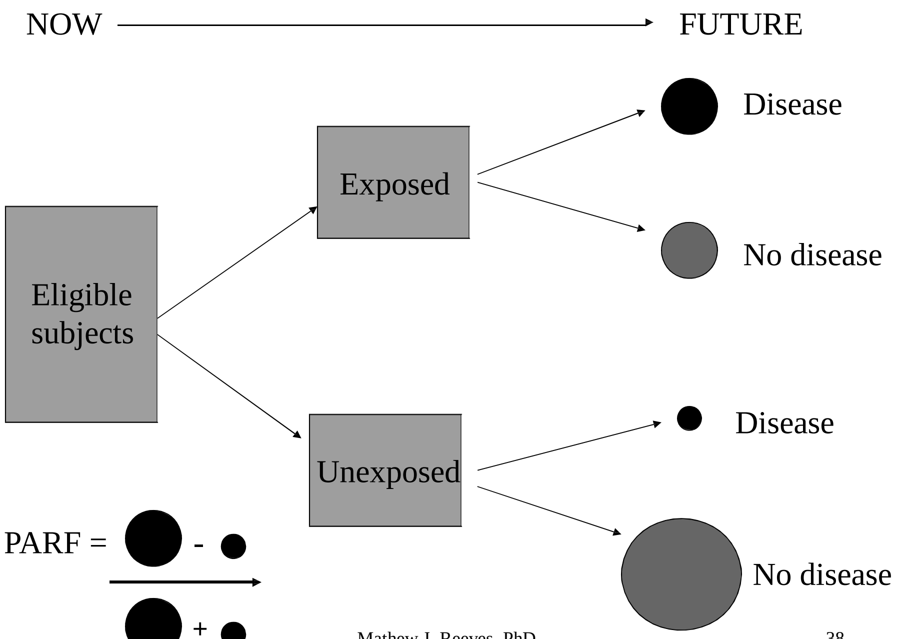
- The incidence of disease in a population that is associated with a risk factor.
- Calculated from the Attributable risk (or RD) and the prevalence (P) of the risk factor in the population
  - PAR = Attributable risk x P
  - PAR = (16-4) x 0.25
  - PAR = 3 per 10,000 person years
- Equals the excess incidence of VTE in the population due to OC use

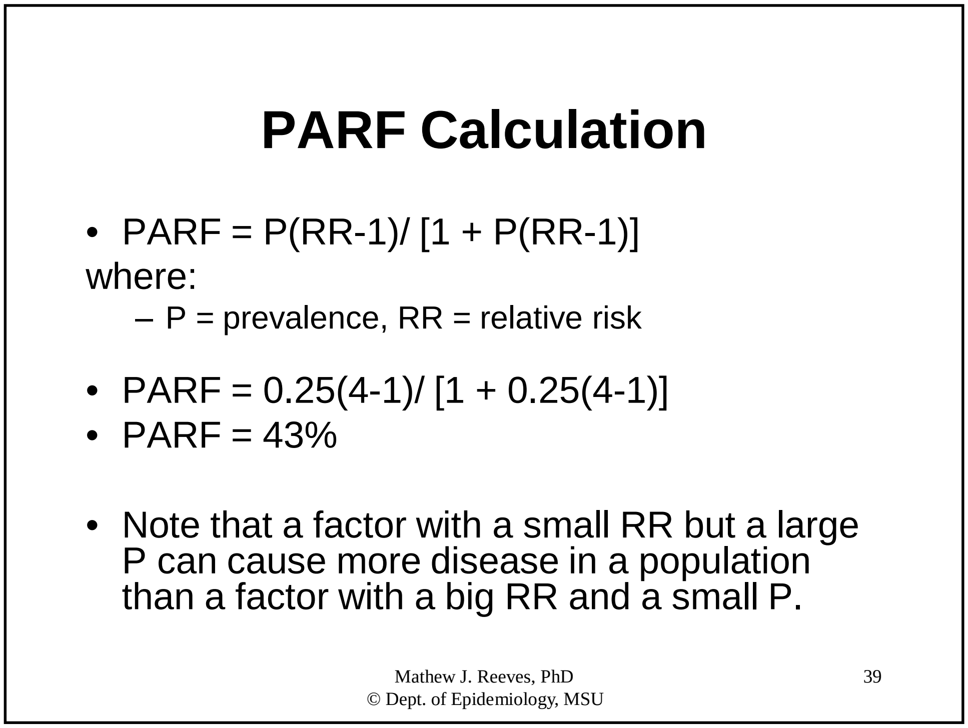
Population attributable risk fraction (PARF)

- The fraction of disease in a population that is attributed to a risk factor.

- PARF = PAR/Total incidence
  - PARF = 3/7per 10,000 person years
  - PARF = 43%

- Represents the maximum potential impact on disease incidence if risk factor was removed
  - So, remove OC's and incidence of VTE drops 43% in women of repro age (assuming OC is a cause of VTE)





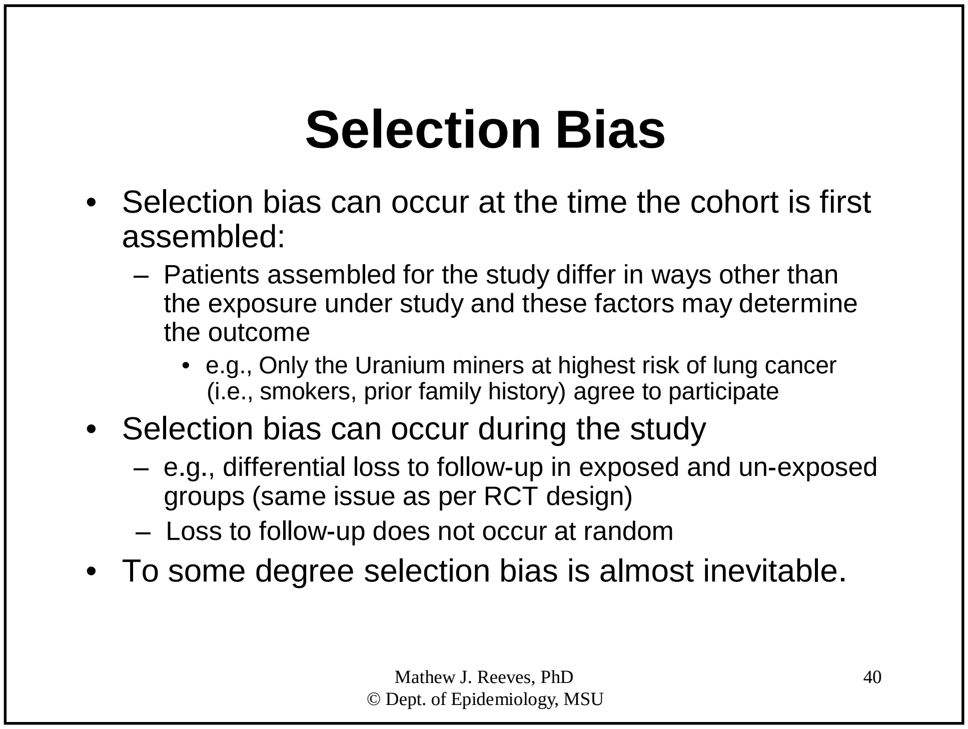
PARF Calculation

- PARF = P(RR-1)/ [1 + P(RR-1)]

where:

  – P = prevalence, RR = relative risk

- PARF = 0.25(4-1)/ [1 + 0.25(4-1)]
- PARF = 43%
- Note that a factor with a small RR but a large P can cause more disease in a population than a factor with a big RR and a small P.

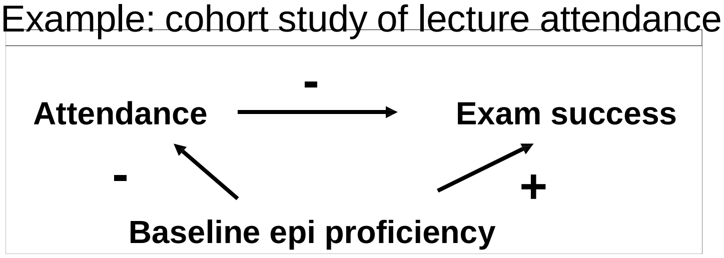


Selection Bias

- Selection bias can occur at the time the cohort is first assembled:
  - Patients assembled for the study differ in ways other than the exposure under study and these factors may determine the outcome
    - e.g., Only the Uranium miners at highest risk of lung cancer (i.e., smokers, prior family history) agree to participate
- Selection bias can occur during the study
  - e.g., differential loss to follow-up in exposed and un-exposed groups (same issue as per RCT design)
  - Loss to follow-up does not occur at random
- To some degree selection bias is almost inevitable.

Confounding Bias

- Confounding bias can occur in cohort studies because the exposure of interest is not assigned at random and other risk factors may be associated with both the exposure and disease.
- Example: cohort study of lecture attendance



Attendance

-

Exam success

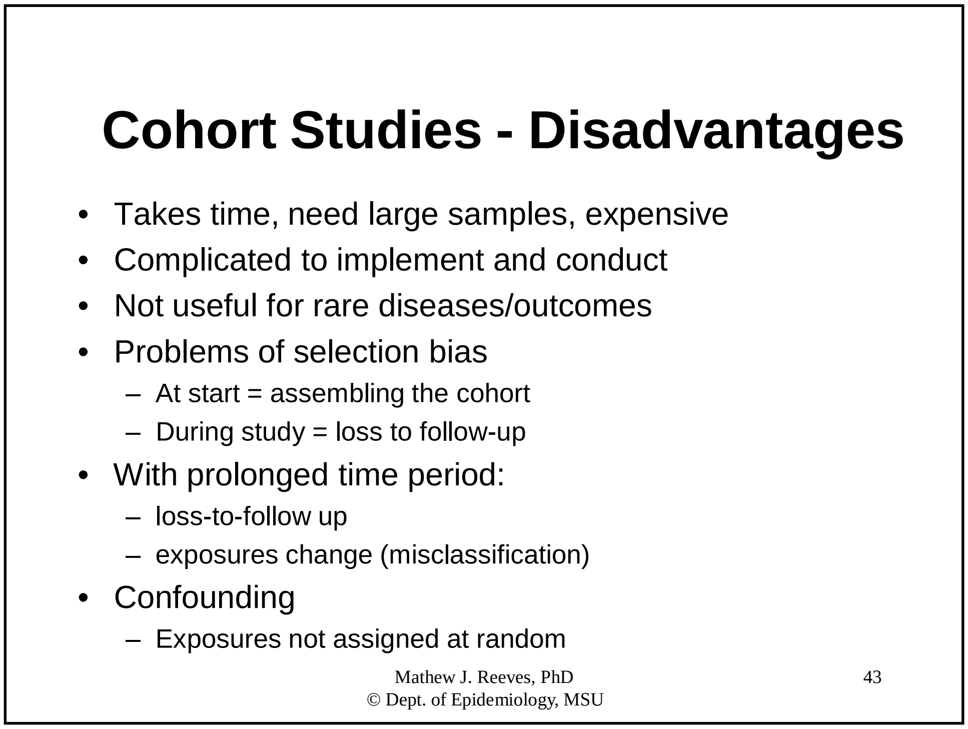
-

+

Baseline epi proficiency

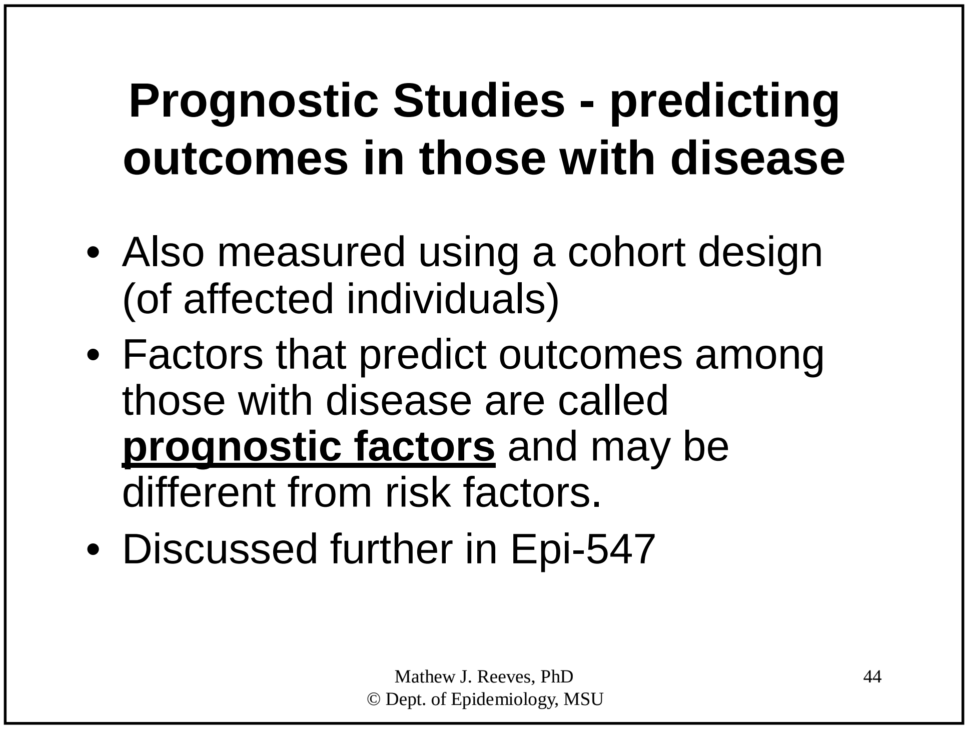
Cohort Studies - Advantages

- Can measure disease incidence
- Can study the natural history
- Provides strong evidence of casual association between E and D (time order is known)
- Provides information on time lag between E and D
- Multiple diseases can be examined
- Good choice if exposure is rare (assemble special exposure cohort)
- Generally less susceptible to bias vs. CCS



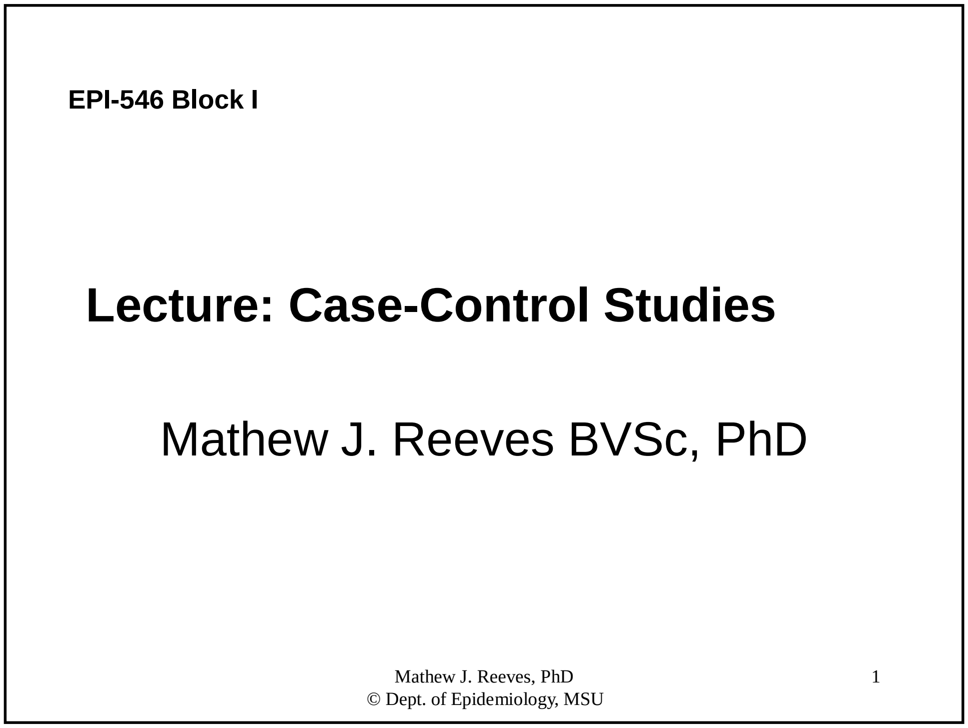
Cohort Studies - Disadvantages

- Takes time, need large samples, expensive
- Complicated to implement and conduct
- Not useful for rare diseases/outcomes
- Problems of selection bias
  - At start = assembling the cohort
  - During study = loss to follow-up
- With prolonged time period:
  - loss-to-follow up
  - exposures change (misclassification)
- Confounding
  - Exposures not assigned at random



Prognostic Studies - predicting outcomes in those with disease

- Also measured using a cohort design (of affected individuals)
- Factors that predict outcomes among those with disease are called **prognostic factors** and may be different from risk factors.
- Discussed further in Epi-547



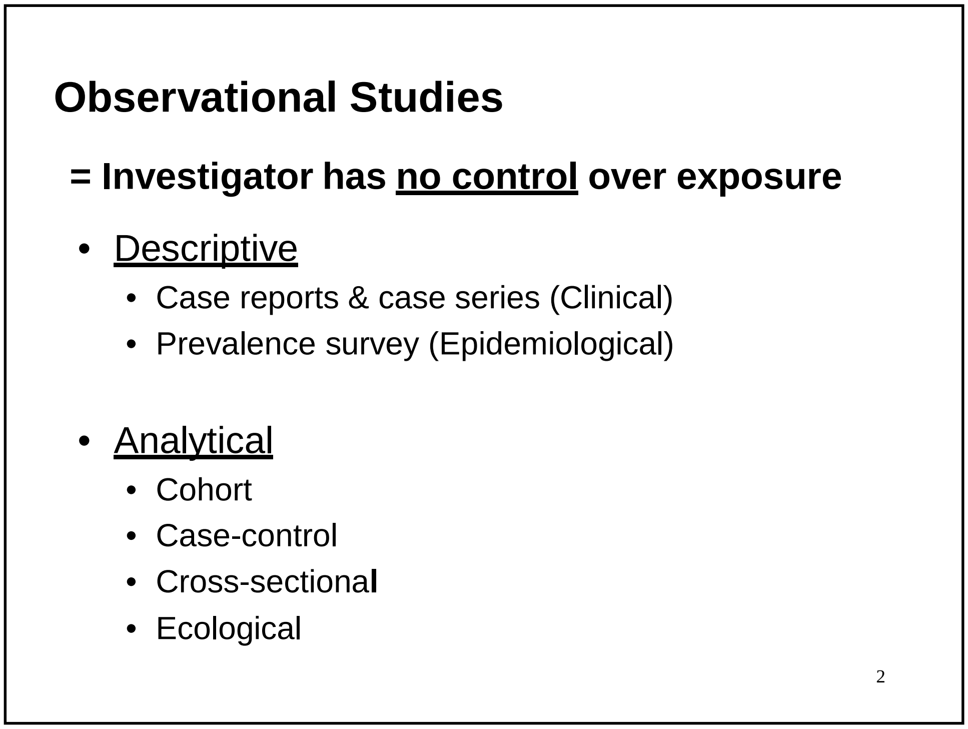
EPI-546 Block I

Lecture: Case-Control Studies

Mathew J. Reeves BVSc, PhD

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

1



Observational Studies

= Investigator has no control over exposure

- Descriptive
  - Case reports & case series (Clinical)
  - Prevalence survey (Epidemiological)
- Analytical
  - Cohort
  - Case-control
  - Cross-sectional
  - Ecological

2

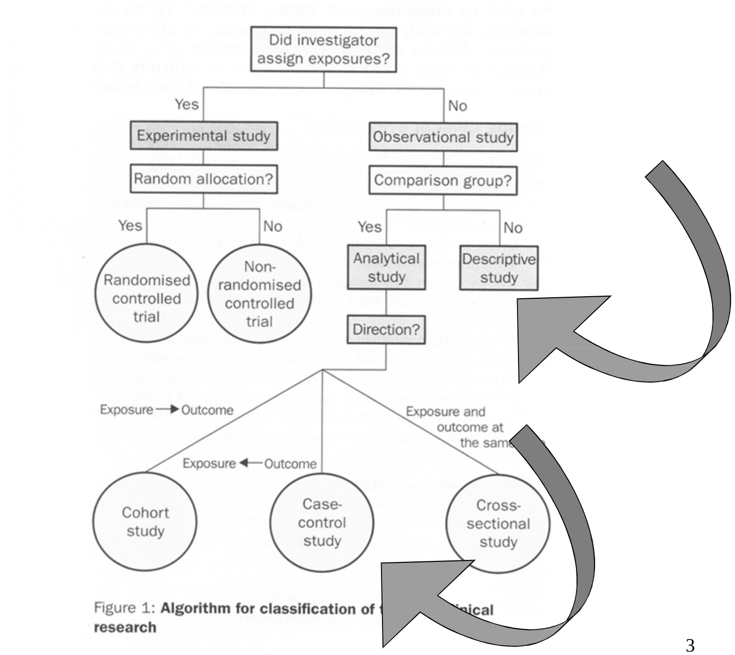


Figure 1: **Algorithm for classification of t inical research**

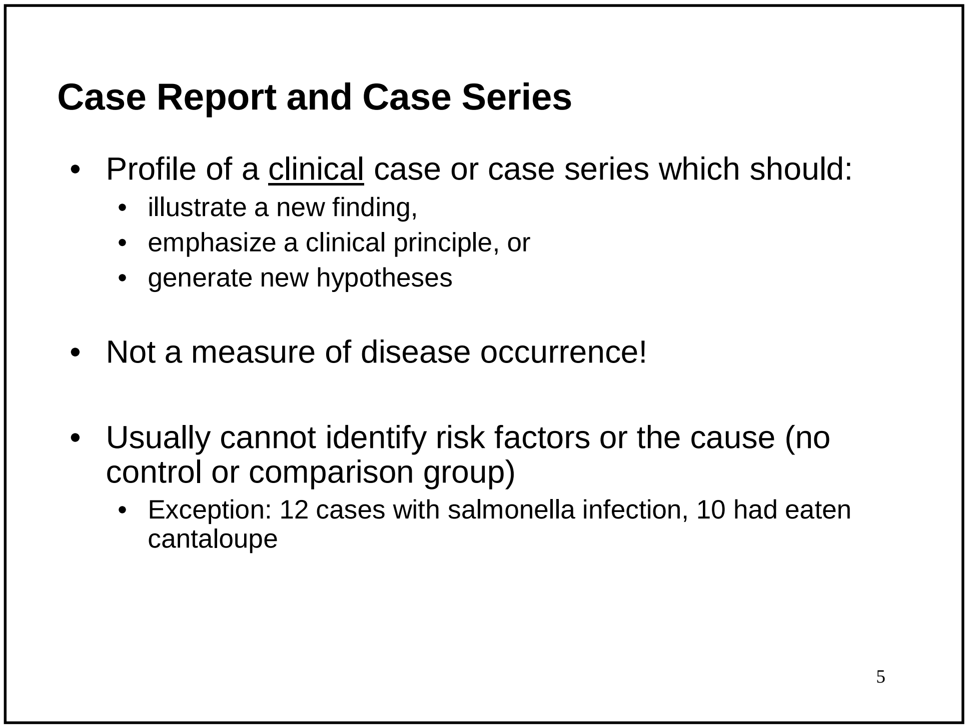
3

Grimes DA and Schulz KF 2002. An overview of clinical research. Lancet 359:57-61.

Objectives – CCS Concepts

- Define and identify case reports and case series
- Define, understand and identify (CCS)
  - Distinguish CCS from other designs (esp. retrospective cohort)
- Understand the principles of selecting cases and controls
- Understand the analysis of CCS
  - Calculation and interpretation of the OR
- Understand the concept of matching
- Understand the origin and consequence of recall bias
  - Example of measurement bias
- Advantages and disadvantages of CCS

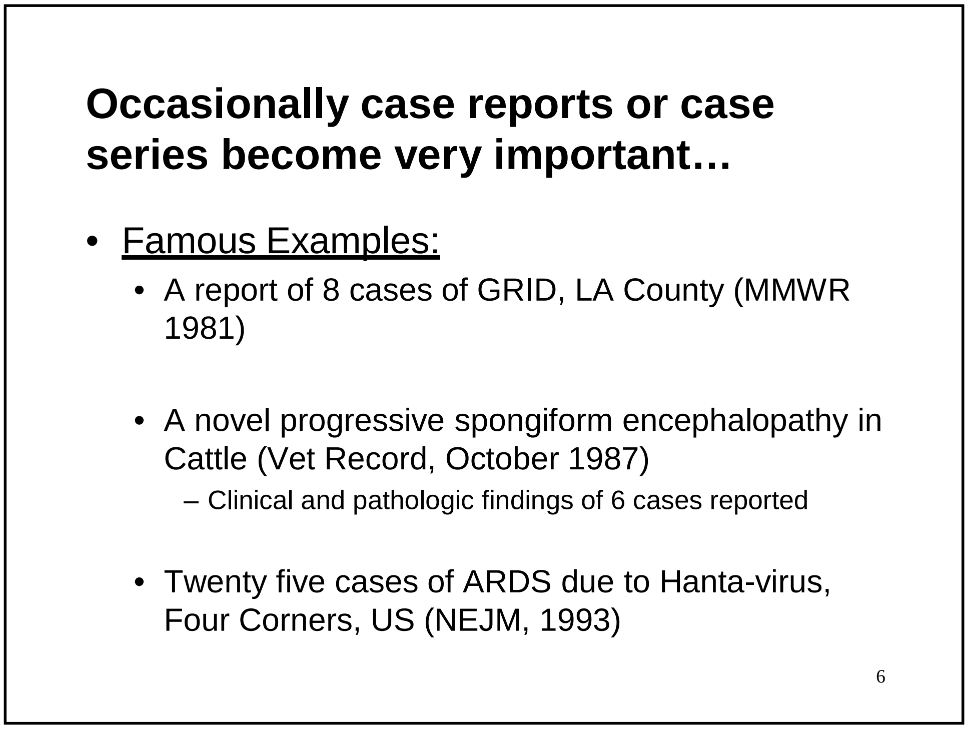
4



Case Report and Case Series

- Profile of a clinical case or case series which should:
  - illustrate a new finding,
  - emphasize a clinical principle, or
  - generate new hypotheses
- Not a measure of disease occurrence!
- Usually cannot identify risk factors or the cause (no control or comparison group)
  - Exception: 12 cases with salmonella infection, 10 had eaten cantaloupe

5



Occasionally case reports or case series become very important…

- Famous Examples:
  - A report of 8 cases of GRID, LA County (MMWR 1981)
  - A novel progressive spongiform encephalopathy in Cattle (Vet Record, October 1987)
    - – Clinical and pathologic findings of 6 cases reported
  - Twenty five cases of ARDS due to Hanta-virus, Four Corners, US (NEJM, 1993)

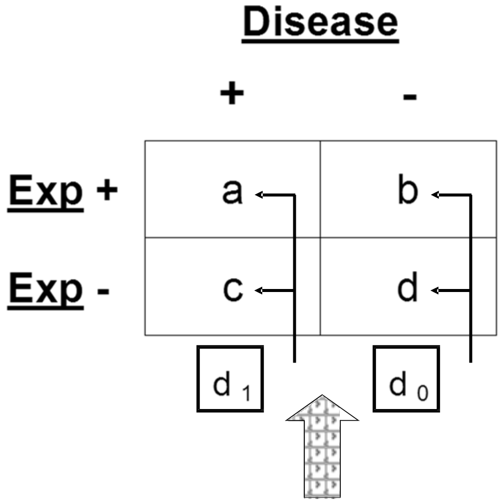
6

Case-Control Studies (CCS)

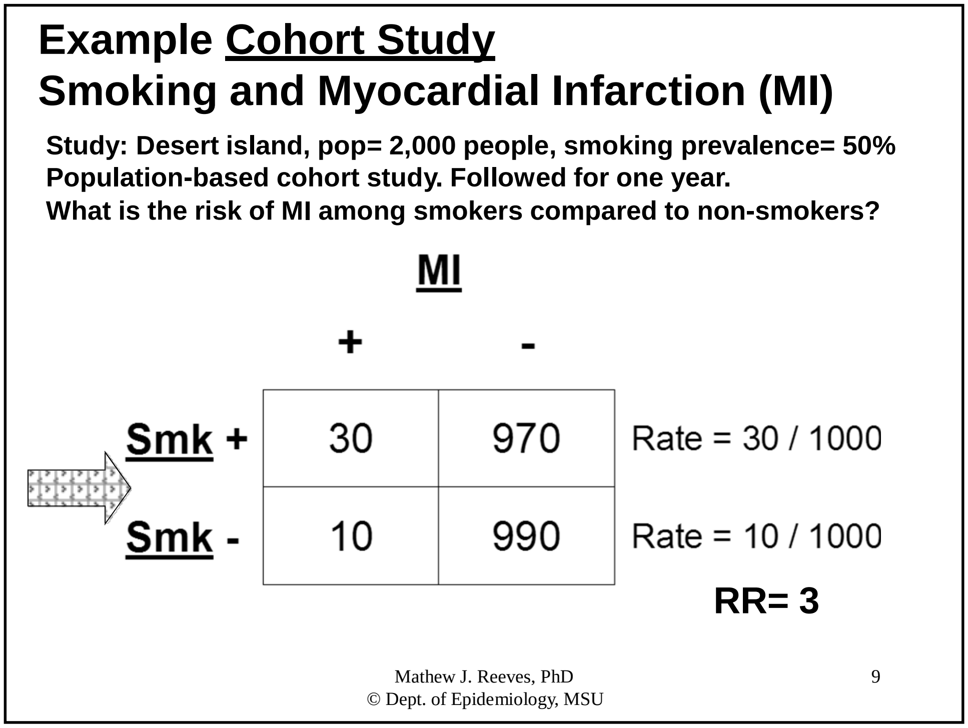
- An alternative observational design to identify risk factors for a disease/outcome.
- Question:
  - How do diseased cases differ from non-diseased (controls) with respect to prior exposure history?
  - Compare frequency of <u>exposure</u> among cases and controls
  - Effect ⟶ cause.
  - <u>Cannot</u> calculate disease <u>incidence</u> rates because the CCS does not follow a disease free- population over time

Case-control Study – Design

Select subjects on the basis of disease status



	Disease +	**Disease** -
Exp +	a	b
Exp -	c	d



Example <u>Cohort Study</u>
Smoking and Myocardial Infarction (MI)

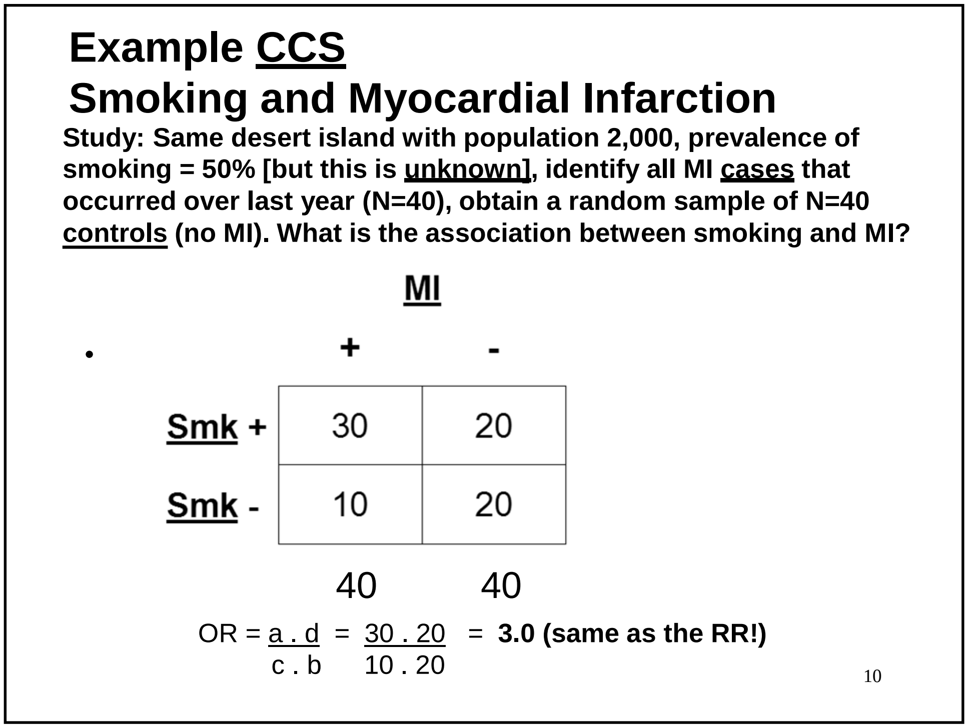
Study: Desert island, pop= 2,000 people, smoking prevalence= 50%
Population-based cohort study. Followed for one year.
What is the risk of MI among smokers compared to non-smokers?

	MI +	MI -	
Smk +	30	970	Rate = 30 / 1000
Smk -	10	990	Rate = 10 / 1000

RR= 3

Mathew J. Reeves, PhD
© Dept. of Epidemiology, MSU

9



Example <u>CCS</u>
Smoking and Myocardial Infarction

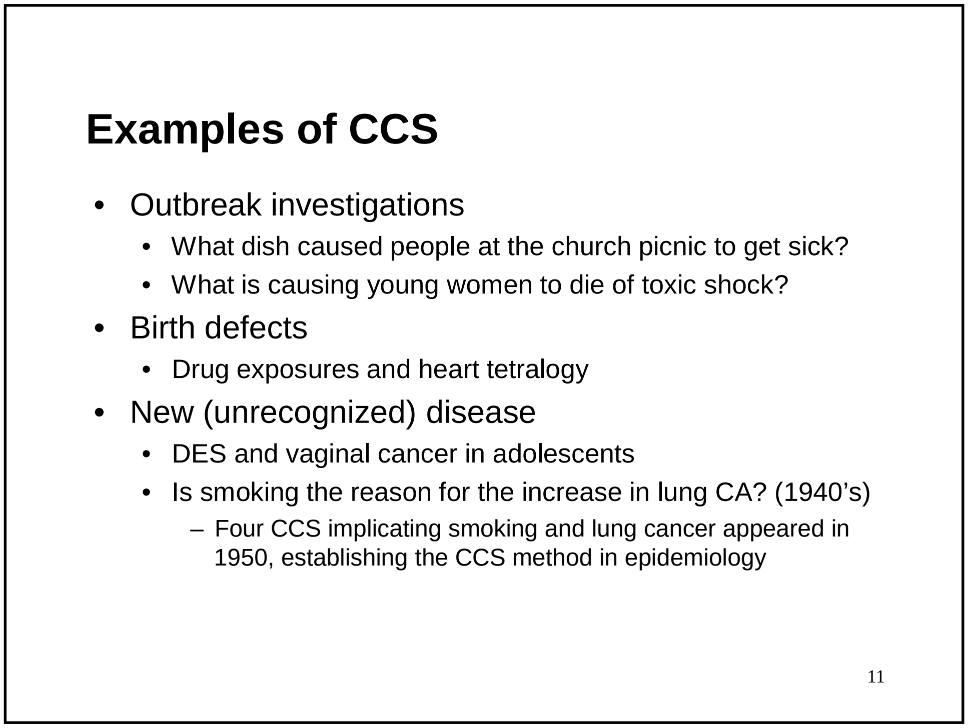
Study: Same desert island with population 2,000, prevalence of smoking = 50% [but this is <u>unknown</u>], identify all MI <u>cases</u> that occurred over last year (N=40), obtain a random sample of N=40 <u>controls</u> (no MI). What is the association between smoking and MI?

•

	MI +	MI -
Smk +	30	20
Smk -	10	20
	40	40

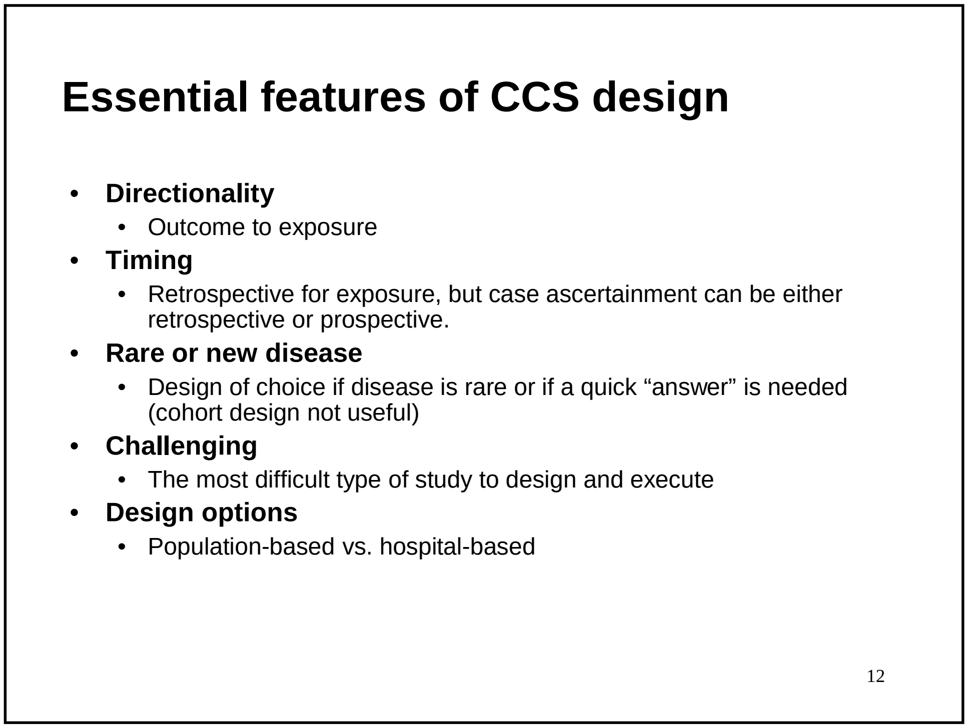
$OR = \frac{a . d}{c . b} = \frac{30 . 20}{10 . 20}$ = **3.0 (same as the RR!)**

10



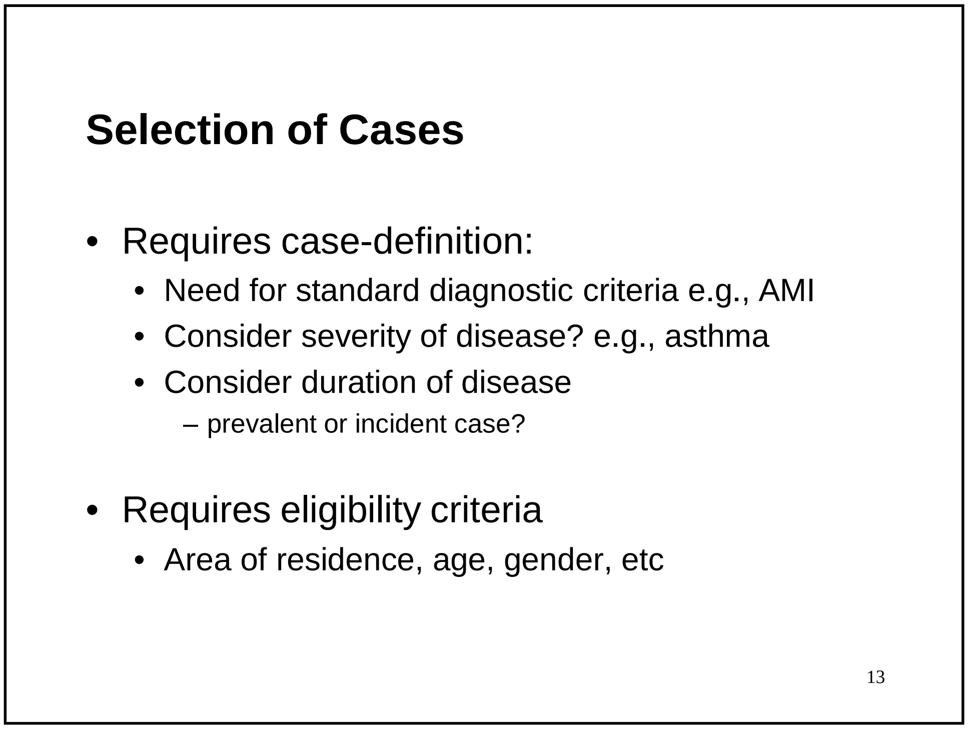
Examples of CCS

- Outbreak investigations
  - What dish caused people at the church picnic to get sick?
  - What is causing young women to die of toxic shock?
- Birth defects
  - Drug exposures and heart tetralogy
- New (unrecognized) disease
  - DES and vaginal cancer in adolescents
  - Is smoking the reason for the increase in lung CA? (1940’s)
    - Four CCS implicating smoking and lung cancer appeared in 1950, establishing the CCS method in epidemiology



Essential features of CCS design

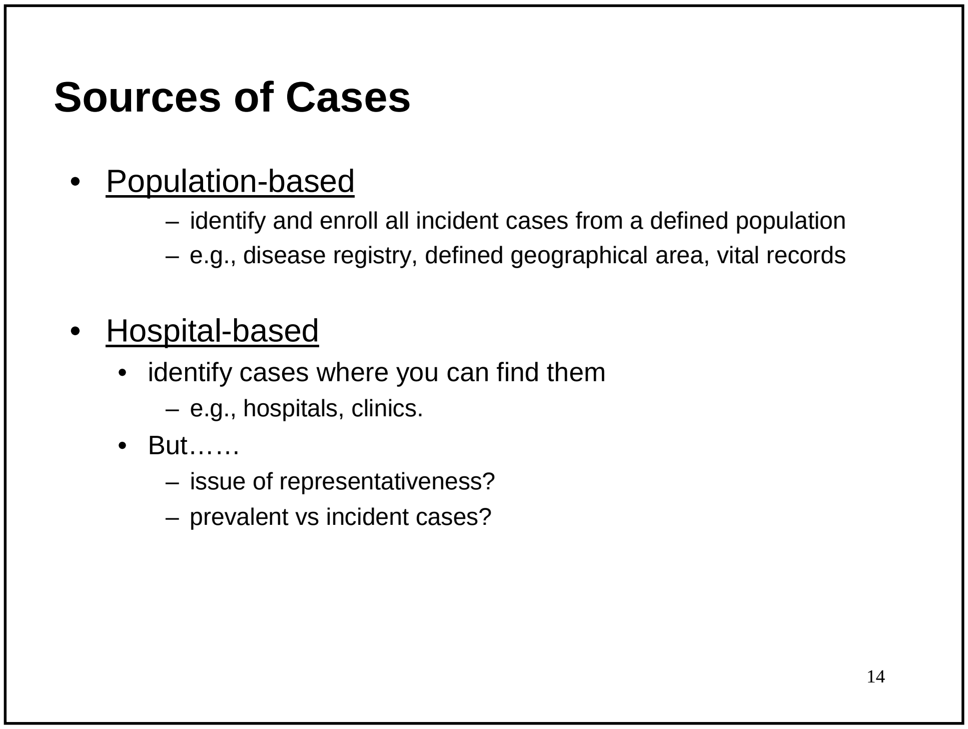
- **Directionality**
  - Outcome to exposure
- **Timing**
  - Retrospective for exposure, but case ascertainment can be either retrospective or prospective.
- **Rare or new disease**
  - Design of choice if disease is rare or if a quick “answer” is needed (cohort design not useful)
- **Challenging**
  - The most difficult type of study to design and execute
- **Design options**
  - Population-based vs. hospital-based



Selection of Cases

- Requires case-definition:
  - Need for standard diagnostic criteria e.g., AMI
  - Consider severity of disease? e.g., asthma
  - Consider duration of disease
    - – prevalent or incident case?
- Requires eligibility criteria
  - Area of residence, age, gender, etc

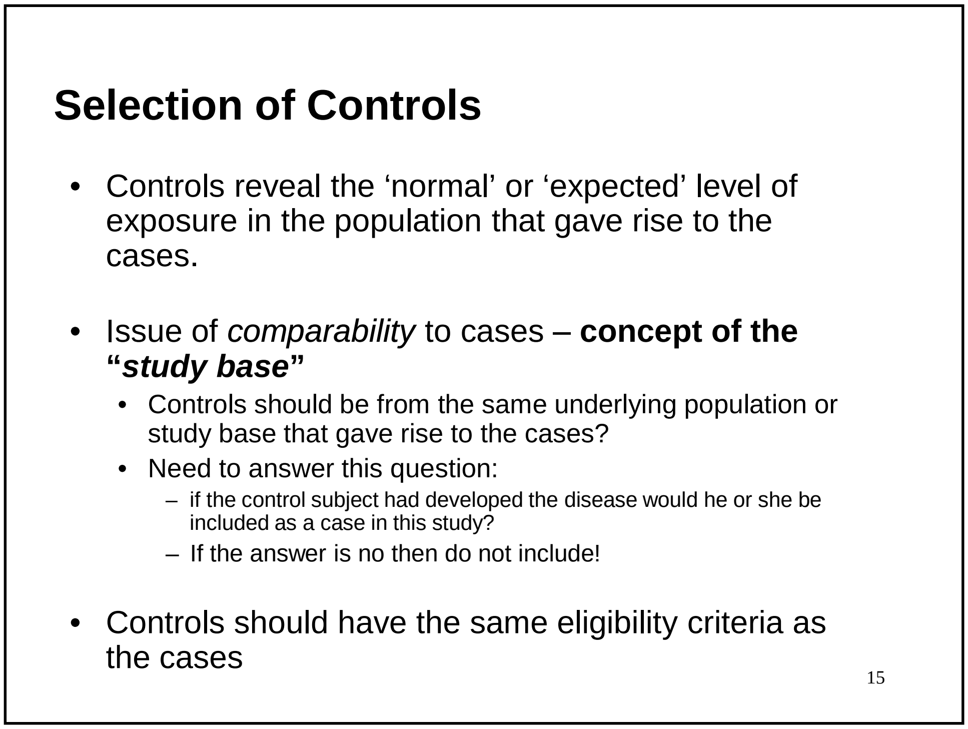
13



Sources of Cases

- Population-based
    - – identify and enroll all incident cases from a defined population
    - – e.g., disease registry, defined geographical area, vital records
- Hospital-based
  - identify cases where you can find them
    - – e.g., hospitals, clinics.
  - But……
    - – issue of representativeness?
    - – prevalent vs incident cases?

14

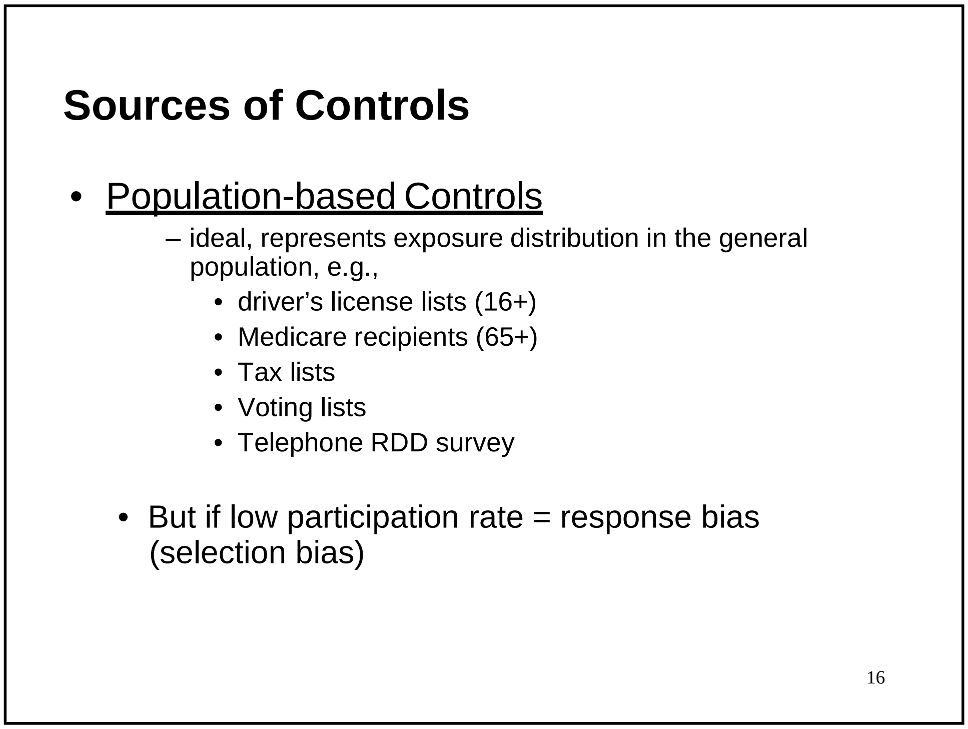


Selection of Controls

- Controls reveal the ‘normal’ or ‘expected’ level of exposure in the population that gave rise to the cases.

- Issue of *comparability* to cases – **concept of the “*study base*”**
  - Controls should be from the same underlying population or study base that gave rise to the cases?
  - Need to answer this question:
    - if the control subject had developed the disease would he or she be included as a case in this study?
    - If the answer is no then do not include!

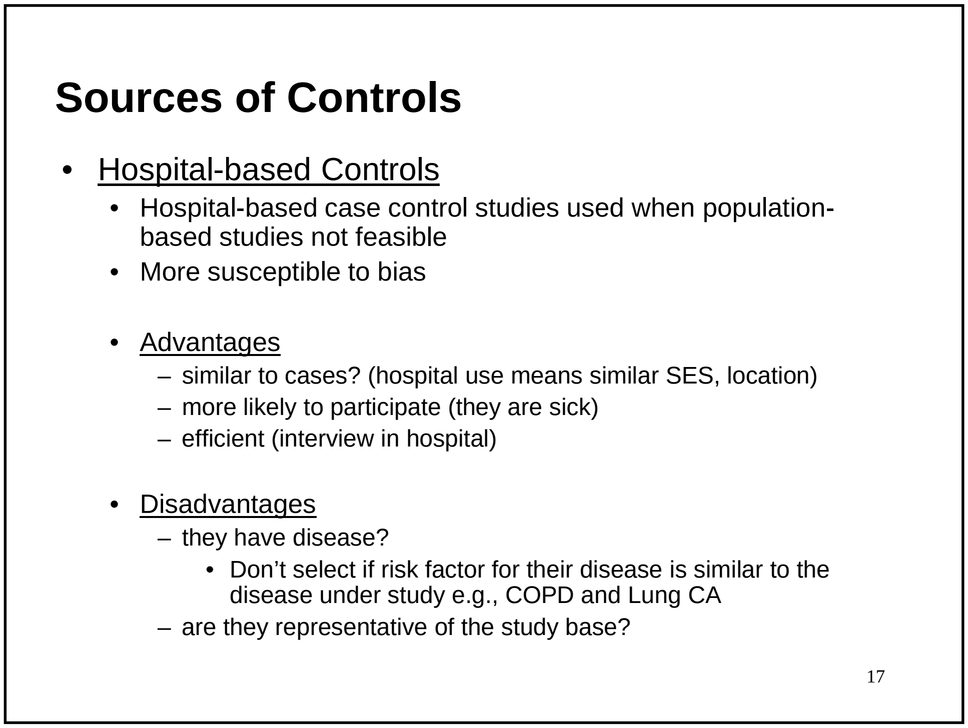
- Controls should have the same eligibility criteria as the cases



Sources of Controls

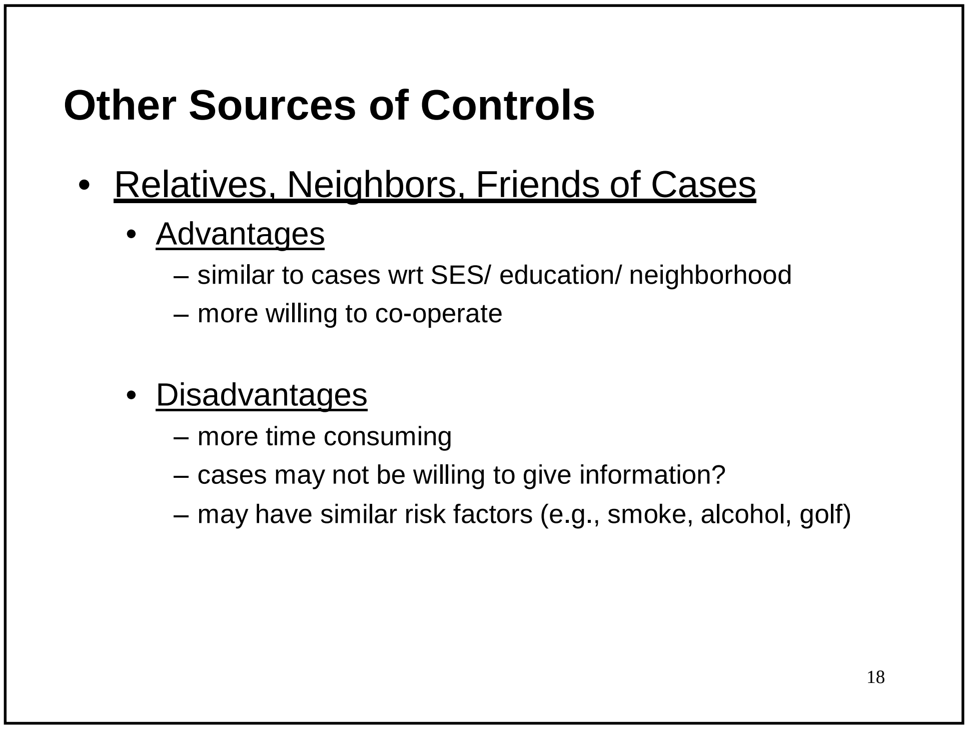
- Population-based Controls
    - ideal, represents exposure distribution in the general population, e.g.,
      - driver’s license lists (16+)
      - Medicare recipients (65+)
      - Tax lists
      - Voting lists
      - Telephone RDD survey

  - But if low participation rate = response bias (selection bias)



Sources of Controls

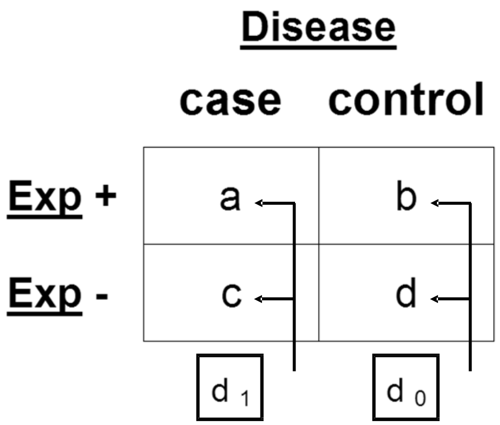
- Hospital-based Controls
  - Hospital-based case control studies used when population-based studies not feasible
  - More susceptible to bias
  - Advantages
    - similar to cases? (hospital use means similar SES, location)
    - more likely to participate (they are sick)
    - efficient (interview in hospital)
  - Disadvantages
    - they have disease?
      - Don’t select if risk factor for their disease is similar to the disease under study e.g., COPD and Lung CA
    - are they representative of the study base?



Other Sources of Controls

- Relatives, Neighbors, Friends of Cases
  - Advantages
    - similar to cases wrt SES/ education/ neighborhood
    - more willing to co-operate
  - Disadvantages
    - more time consuming
    - cases may not be willing to give information?
    - may have similar risk factors (e.g., smoke, alcohol, golf)

- **Analysis of CCS**
  - Odds of exposure among cases = a / c
  - Odds of exposure among controls = b / d



	Disease case	 control
Exp +	a	b
Exp -	c	d
	d_1	d_0

Analysis of CCS

The OR is the only measure of association

- The only valid measure of association for the CCS is the *Odds Ratio* (OR)

- Under reasonable assumptions (– the rare disease assumption) the **OR approximates the RR**.

- OR = $\frac{\text{Odds of exposure among cases (disease)}}{\text{Odds of exposure among controls (non-dis)}}$

  - Odds of exposure among cases = a / c
  - Odds of exposure among controls = b / d
  - Odds ratio = $\frac{a/c}{b/d} = \frac{a.d}{b.c}$ [= cross-product ratio]

Example CCS
Smoking and Myocardial Infarction

Study: Same desert island with population 2,000, prevalence of smoking = 50% [but this is unknown], identify all MI cases that occurred over last year (N=40), obtain a random sample of N=40 controls (no MI). What is the association between smoking and MI?

	MI +	MI -
Smk +	30	20
Smk -	10	20
	40	40

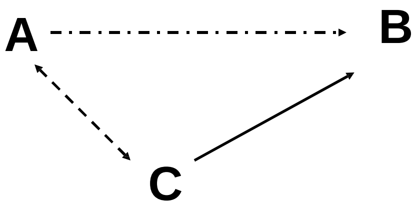
$$OR = \frac{a \, . \, d}{c \, . \, b} = \frac{30 \, . \, 20}{10 \, . \, 20} = \mathbf{3.0}$$ **(same as the RR!)**

Odds Ratio (OR)

- Similar interpretation as the Relative Risk
- OR = 1.0 (implies equal odds of exposure - no effect)
- ORs provide the exact same information as the RR if:
  - controls represent the target population
  - cases represent all cases
  - rare disease assumption holds (or if case-control study is undertaken with population-based sampling)
- Remember:
  - OR can be calculated for any design but RR can only be calculated in RCT and cohort studies
  - The OR is the only valid measure for CCS
  - Publications will occasionally mis-label OR as RR (or vice versa)

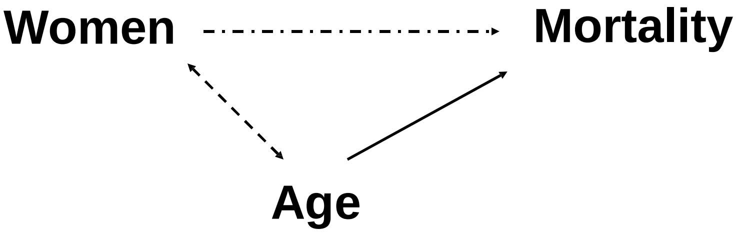
Controlling extraneous variables (confounding)

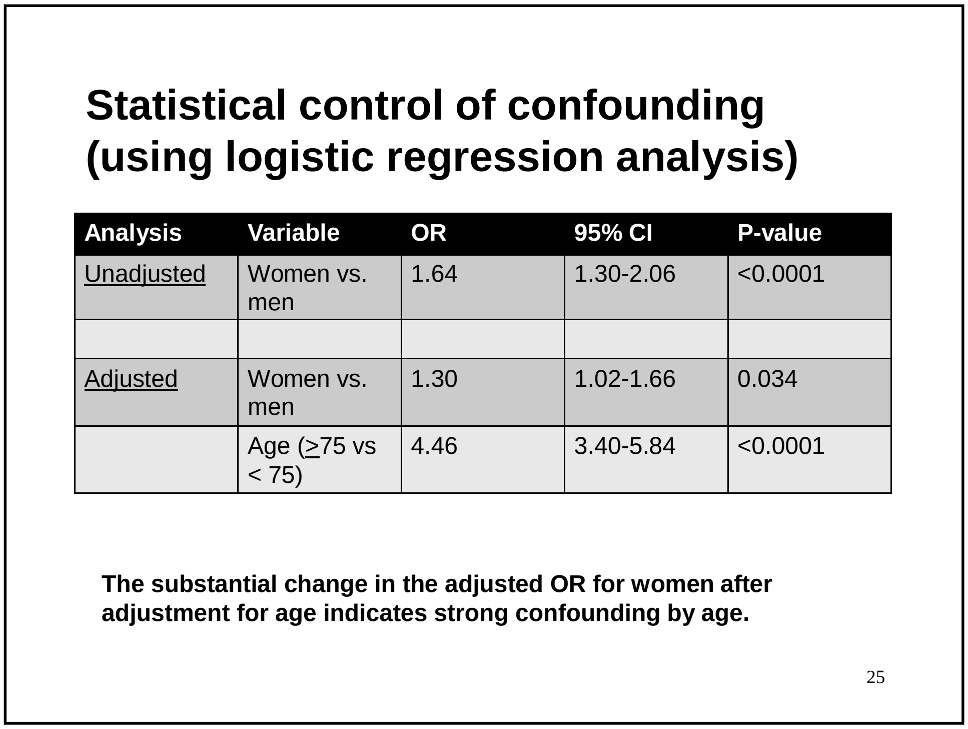
- Exposure of interest may be confounded by a factor that is associated with the exposure **and** the disease i.e., is an independent risk factor for the disease



Example – Sex differences in stroke survival and confounding by age

- Women have higher risk of mortality following stroke vs. men (35% vs. 24%; OR = 1.6, 95% CI 1.29-2.06)
- But women are older than men (77 vs 72 years)
- Age is an important risk factor for dying



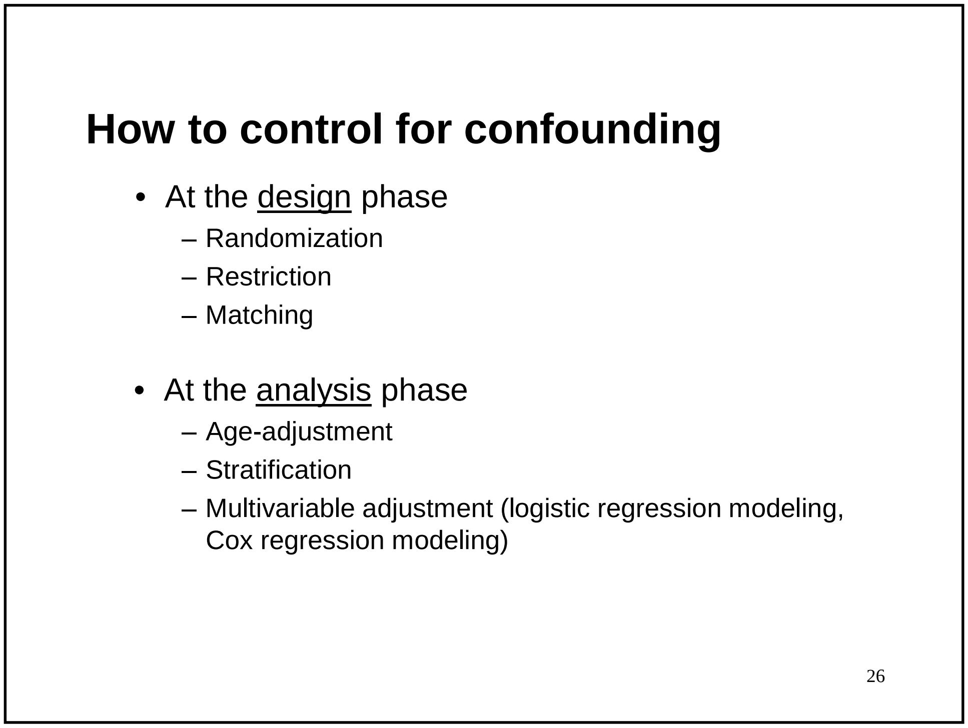


Statistical control of confounding (using logistic regression analysis)

Analysis	Variable	OR	95% CI	P-value
Unadjusted	Women vs. men	1.64	1.30-2.06	<0.0001
Adjusted	Women vs. men	1.30	1.02-1.66	0.034
	Age ($\geq$75 vs < 75)	4.46	3.40-5.84	<0.0001

The substantial change in the adjusted OR for women after adjustment for age indicates strong confounding by age.

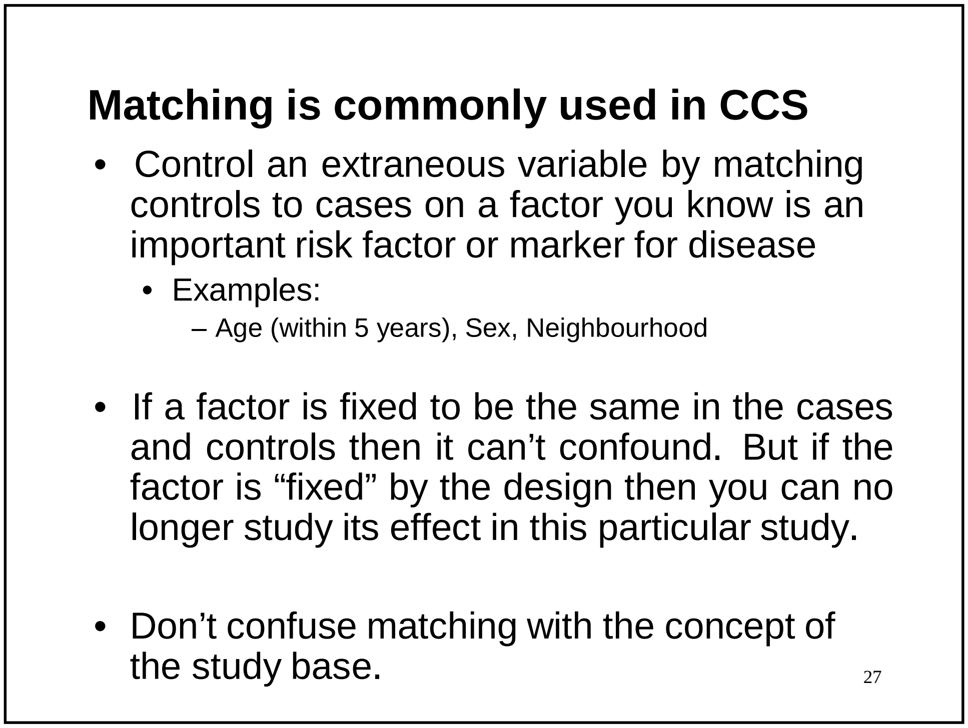
25



How to control for confounding

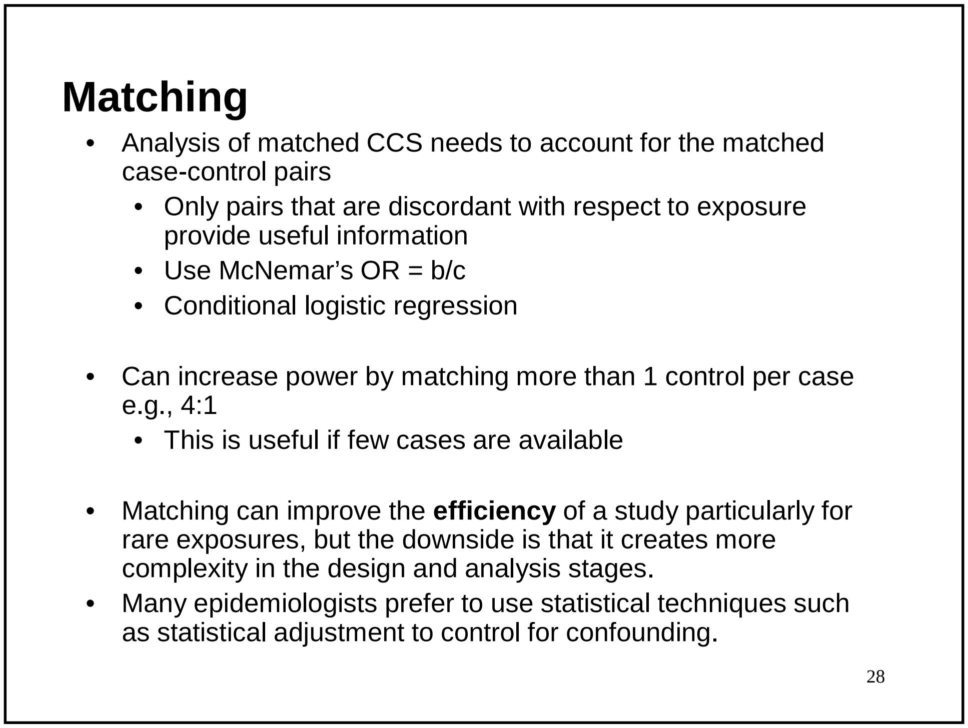
- At the design phase
  - Randomization
  - Restriction
  - Matching
- At the analysis phase
  - Age-adjustment
  - Stratification
  - Multivariable adjustment (logistic regression modeling, Cox regression modeling)

26



Matching is commonly used in CCS

- Control an extraneous variable by matching controls to cases on a factor you know is an important risk factor or marker for disease
  - Examples:
    - Age (within 5 years), Sex, Neighbourhood
- If a factor is fixed to be the same in the cases and controls then it can’t confound. But if the factor is “fixed” by the design then you can no longer study its effect in this particular study.
- Don’t confuse matching with the concept of the study base.



Matching

- Analysis of matched CCS needs to account for the matched case-control pairs
  - Only pairs that are discordant with respect to exposure provide useful information
  - Use McNemar’s OR = b/c
  - Conditional logistic regression
- Can increase power by matching more than 1 control per case e.g., 4:1
  - This is useful if few cases are available
- Matching can improve the **efficiency** of a study particularly for rare exposures, but the downside is that it creates more complexity in the design and analysis stages.
- Many epidemiologists prefer to use statistical techniques such as statistical adjustment to control for confounding.

Matched CCS - Discordant pairs

Match 40 controls to 40 cases of AMI so they have the same age and sex. Then classify according to smoking status. Only the discordant cells (‘b’ and ‘c’) contribute to the OR.

•

		Controls	
		+	-
Cases	+	32	20
	-	10	18

80

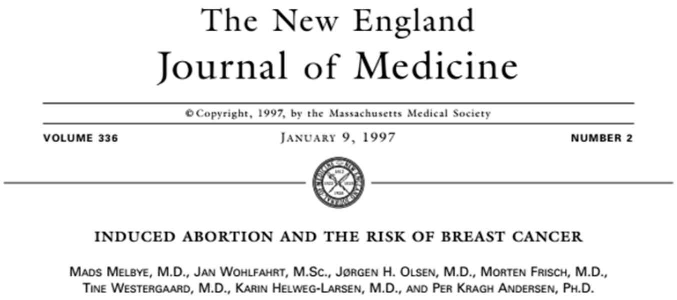
McNemar’s OR = $\frac{b}{c} = \frac{20}{10} = 2.0$

Over-matching

- Matching can result in controls being so similar to cases that all of the exposures are the same
- Example:
  - 8 cases of GRID, LA County, 1981
  - All cases are gay men so match with other gay men who did not have signs of GRID
  - Use 4:1 matching ration i.e. 32 controls
  - No differences found in sexual or other lifestyle habits

Recall Bias

- Form of <u>measurement bias.</u>
- Presence of disease may affect ability to recall or report the exposure.
- Example – exposure to OTC drugs during pregnancy use by moms of normal and congenitally abnormal babies.
- To lessen potential:
  - Blind participants to study hypothesis
  - Blind study personnel to hypothesis
  - Use explicit definitions for exposure
  - Use controls with an unrelated but similar disease
    - E.g., heart tetralogy (cases), hypospadia (controls)



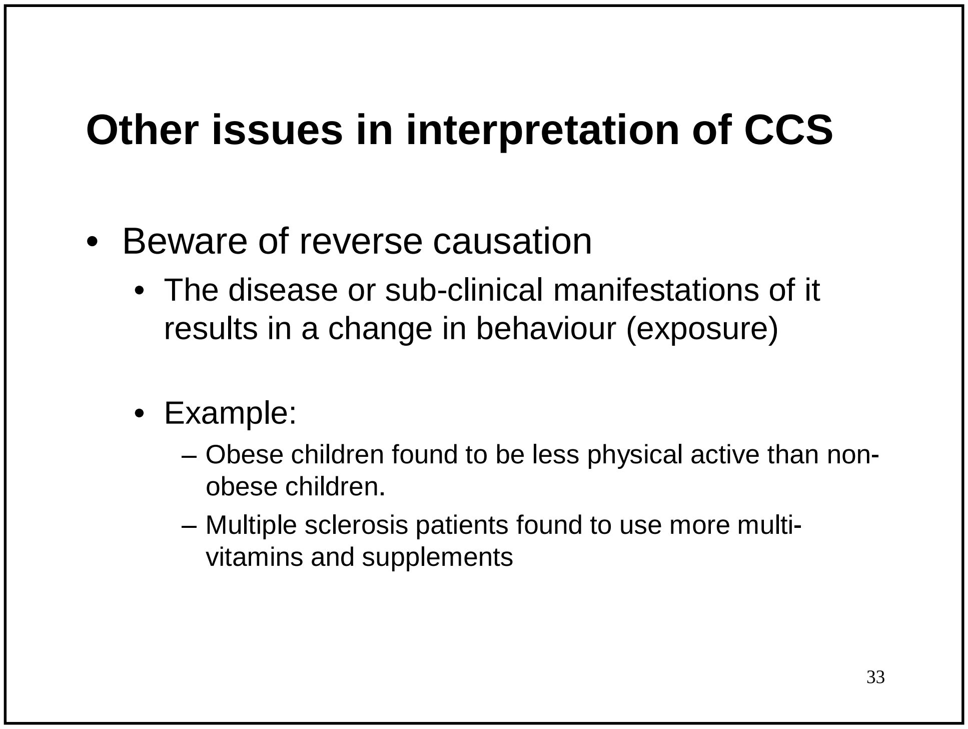
The New England
Journal of Medicine

©Copyright, 1997, by the Massachusetts Medical Society

VOLUME 336 JANUARY 9, 1997 NUMBER 2

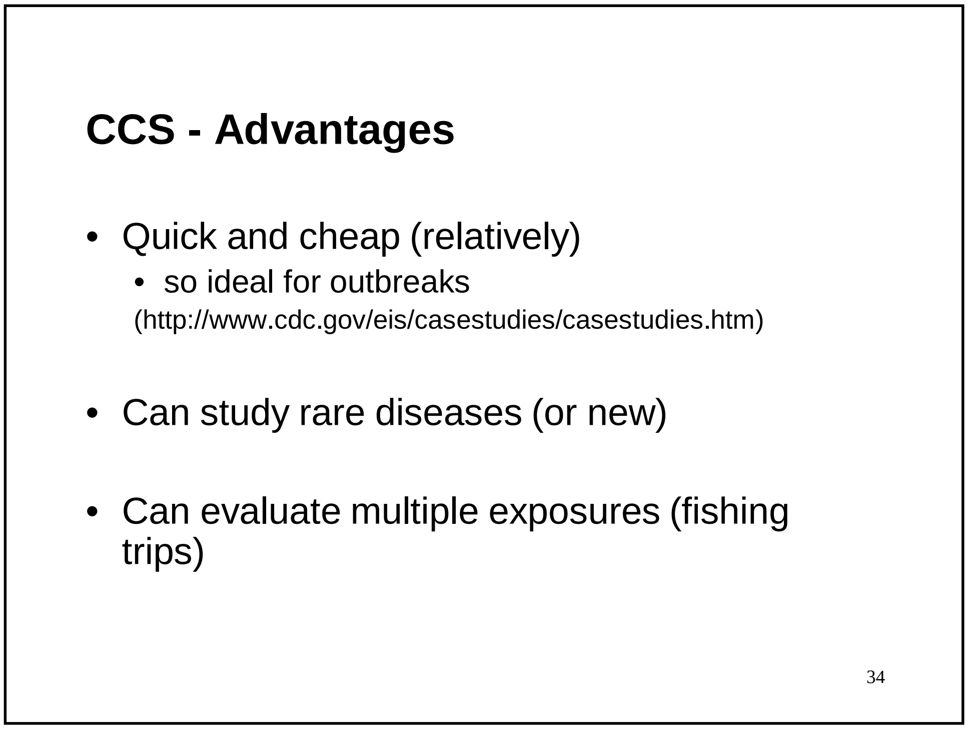
INDUCED ABORTION AND THE RISK OF BREAST CANCER

MADS MELBYE, M.D., JAN WOHLFAHRT, M.SC., JØRGEN H. OLSEN, M.D., MORTEN FRISCH, M.D.,
TINE WESTERGAARD, M.D., KARIN HELWEG-LARSEN, M.D., AND PER KRAGH ANDERSEN, PH.D.



Other issues in interpretation of CCS

- Beware of reverse causation
  - The disease or sub-clinical manifestations of it results in a change in behaviour (exposure)
  - Example:
    - Obese children found to be less physical active than non-obese children.
    - Multiple sclerosis patients found to use more multi-vitamins and supplements

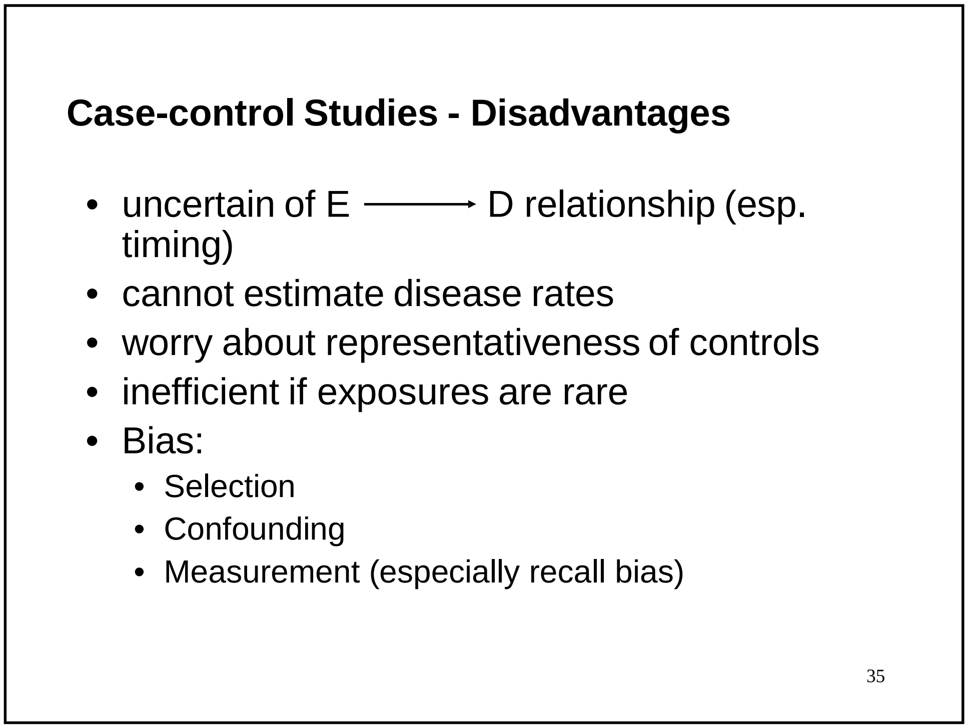


CCS - Advantages

- Quick and cheap (relatively)
  - so ideal for outbreaks

(http://www.cdc.gov/eis/casestudies/casestudies.htm)

- Can study rare diseases (or new)
- Can evaluate multiple exposures (fishing trips)



Case-control Studies - Disadvantages

- uncertain of E → D relationship (esp. timing)
- cannot estimate disease rates
- worry about representativeness of controls
- inefficient if exposures are rare
- Bias:
  - Selection
  - Confounding
  - Measurement (especially recall bias)