Biostatistics Class Notes

Cedars/Sinai Medical Center – Fall

Jeffrey Gornbein, Dr PH
Life Science 5202 – UCLA
621 Charles Young Dr South
310-825-4193
gornbein@g.ucla.edu

Course website: http://gornbein.bol.ucla.edu/cedarassign.htm

CLASS NOTES (Gornbein) - Introduction

Introduction

There is a vast medical literature constantly reporting new claims about the causes of disease and the value of treatments. The goal of this course is to help you become a more knowledgeable user of this literature by acquainting you with the statistical methods and arguments employed. Knowledge of study design and statistical methods helps one to judge the validity and quality of a study and distinguish conclusions based on strong evidence from claims based on weak, inconsistent or inconclusive evidence. This is also the start of understanding methods needed to present data from your own research.

Familiarity with statistical methods helps you to defend yourself against those who try to give a false impression of being "scientific" with numbers.

Is there a link between exercise and breast cancer?
Does tonsillectomy reduce respiratory disease?
Is DDI more effective than AZT in slowing the progress of AIDS?

These are the sorts of questions that are addressed in the medical literature. Proper research design and use of statistics give better answers to these questions.

Class administrative details

1. There are graded homework assignments. Homework is usually due one week after it is assigned or as announced.

2. All of the basic concepts you are responsible for are in these notes. However, some specific examples may not be included and computational details are usually omitted. Such details are in the recommended texts and can be expanded upon during office hours. The class size is small so that we can talk.

3. If you are stuck on an assignment, **ask for help**. You should not be spending more than about 3 hours per assignment. If you find yourself doing a lot more, contact the instructor

3. Office hours are **by appointment**. **Make an appointment!!!**
Instructor- Jeff Gornbein Room Life Science (LS) 5202 - UCLA
office phone: (310) 825-4193
message: (310) 206-1704
fax: (310) 825-8685
electronic mail address: gornbein@ucla.edu

4. General strategy - Memorization and cramming are not very effective in this course. In particular, memorizing formulas without insight is not very useful. The general concepts must "grow" on you. Often, understand the material in one lecture depends on understanding the material in previous lectures. Allow yourself to follow over time and not cram at the end.

Contents (subject to change)

Statistics Notation Summary

Quantity	Population parameter	Sample statistic	usual null value
Mean*	μ	$\overline{X}$	
Mean difference*	$\mu_1 - \mu_2$	$\overline{X_1} - \overline{X_2}$	0
Proportion*	π	P	
Difference in proportions* =Risk difference	$\pi_1 - \pi_2$	$RD=P_1 - P_2$	0
Standard deviation	σ	S	
Correlation coefficient	ρ	r	0
Regression slope	β	b	0
Regression intercept	α	a	0
Risk ratio=relative risk	*RR*	RR	1
Odds ratio	OR	OR	1
Number needed to treat	*NNT*	NNT	not applicable

* Formula for standard error, confidence interval and hypothesis test p value is in the notes

Section I

Logical ideas in medical research

Study Design

Bias and Confounding

Adjusting Rates

I- STUDY DESIGN - Association, causation and comparability
Confounding and bias

Establishing association - the role of design and statistics

In reading the medical literature, much of our interest centers around two types of questions:

1. What is the efficacy of this therapy - does it work? Does is work better than other therapies (including doing nothing).
2. What is the cause of this disease? (Epidemiology)

In short, we are interested in "associations". Associations between possible causes (prognostic factors) and disease **outcomes**, associations between treatments (therapies) and health **outcomes**.

For example:
Is a lumpectomy better than a radical mastectomy? (In terms of a survival outcome)
Does aluminum exposure cause Alzheimers disease?
Does administration of vitamin A reduce measles mortality?

Is there a dose response or exposure response relationship? The higher the dose of an antibiotic, the higher the number of bacteria that are killed. The lower the level of salt, the lower the incidence of coronary heart disease.

Note that the first step is to identify what are the **predictors** (or causes) and what are the **outcomes** or endpoints.

In general, a statistical association between a **predictor** and an **outcome** means that a change in the predictor leads to a change in the outcome. Some persons use the word correlation to mean the same thing as a (statistical) association. In statistics, the word correlation has more restricted, technical definitions and is only a special case of an association.

But just observing a statistical association is not necessarily sufficient to establish **causation** or efficacy.

Superficial association between a predictor and outcome is more likely to be casual if there is (are) no third factor(s) related to both the predictor and outcome (no confounders) and no (internal) bias on the part of the investigator.

Example: AIDS and "amyl nitrate" (poppers). In early 1980s, investigators noted an association between persons using poppers and coming down with AIDS. (for example, see Marmor 1982) Is this proof that AIDS is caused by poppers? (or taking poppers is caused by AIDS?) Notice that this reported association was obtained retrospectively, that is, persons with and without AIDS were questioned about a long list of lifestyle habits. Those with AIDS more frequently reported popper use than those without AIDS. In hindsight we now can see that those who used poppers may have been more likely to have sex with multiple partners and might also be more likely to use drugs intravenously. They therefore had a higher risk of exposure to the HIV virus. So, a third factor, (a **confounding** factor) exposure to a virus, was associated with both poppers and AIDS.

Working definition of causality (or efficacy) - The requirement for "proof"

Definition: We say that "X causes Y" when, all other factors associated with the outcome held constant, a change in predictor X, the "cause" (more frequently) leads to a change in the outcome (or effect) Y. This usually implies a temporal ordering (the cause must happen before the effect) and/or a dose response (the higher the dose of ionizing radiation the higher the probability of getting cancer. So, to establish causality (for disease) or efficacy (for a treatment) there are at least four requirements:

I. Changes in "X" are associated with changes in "Y"

II. The temporal ordering must be correct (cause X comes before effect Y). This is especially challenging in observational studies

III. The association between X and Y must not be due to chance alone. This is where inferential statistics (p values, Cis) are useful.

IV. All other effects on Y that are associated with X must be controlled. For comparing X=groups, this implies that the comparison groups must be comparable (no bias, no confounding). This will not happen unless the study had the proper design.

Bradford Hill "Causastion" Criteria

1. Consistency: Same finding observed by different persons in different places with different samples

2. Specificity: Causation is likely if seen in a very specific population at a specific site and disease with no other likely explanation. The more specific an association between a factor and an effect is, the bigger the probability of a causal relationship.

3. Temporality (III above): The effect has to occur after the cause. If there is an expected delay between the cause and expected effect, then the effect must occur after that delay.

4. Biological gradient: Greater exposure should generally lead to greater incidence. However, in some cases, the mere presence of the factor can trigger the effect. In other cases, an inverse proportion is observed: greater exposure leads to lower incidence. Sometimes called the "dose-response" effect. Can be "U" shaped.

5. Plausibility: A plausible mechanism between cause and effect is helpful, but not required.

6. Coherence: There is coherence (agreement) between epidemiological and laboratory findings .

7. Experiment: Relationship can be investigated in an experiment. Not always possible.

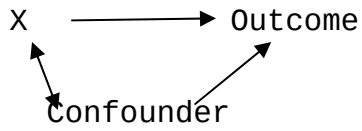
8. Analogy: The effect of similar factors may be considered.

Reasons for lack of comparability -<u>Confounders</u> (uncontrolled prognostic factors)

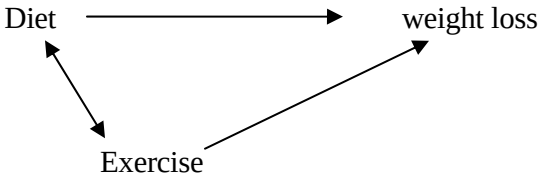
When studying the relationship between a predictor and an outcome, a major goal is to try to hold all other factors constant. At the end of a study, we would like to be able to say that "all else being equal", the relationship between the predictor and outcome is revealed. The problem is in making "all else equal", that is, in controlling for the other "confounding" factors that may distort the relationship we are studying.

For example, when comparing radical mastectomy to "lumpectomy" in the treatment of breast cancer, one important outcome is time until death (also called survival time). In evaluating which surgery allows patients to survive the longest, **<u>age</u>** is a major confounder that must be considered since older women will die sooner than younger women regardless of which therapy they receive. **In general, a confounder is a patient factor or characteristic that is different in each treatment group and influences the study outcome independent of the treatment. (Or independent of the factor under investigation). More generally, a confounder is an independent risk factor that is associated with the factor being studied.** When curing disease is the outcome, initial disease severity is often the most important confounder.

Generally there are two approaches for making all else equal and reducing confounding. One is to **design** the study so that extraneous factors are controlled for or "balance out" in the groups being compared. The other way (which may not be as satisfactory) is to "adjust" for extraneous factors mathematically using stratification, matching or statistical models. **Stratification**, involves looking at the relation between A and B for each subgroup with a constant level of confounding factor C. For example, when evaluating the Pritikin diet versus a low fat, high carbohydrate diet to see which better promotes weight loss, exercise may be a confounding factor. Therefore a separate analysis might be done on those with low, medium and high exercise levels. This would be termed an analysis **<u>stratified</u>** by exercise.

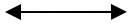


Example: Exercise may be a confounder when determining whether one diet is better than other if exercise itself can affect weight loss and diet.



Key:


= "causation" (unidirectional effect)


= "association"

When disease improvement is the outcome, baseline disease severity and age are commonly important confounding factors**.**

Variables that are NOT confounders – Intermediate or exogenous risk factor
Example: cholesterol level

Does a new drug (B) lower rates of heart attack and stroke compared to the standard drug (A) ?

In this study one would NOT control for the levels of post drug cholesterol level. It is not a confounder but an intermediate risk factor (a "**mediator"** in psychology). If we incorrectly control for cholesterol level, we may reduce the apparent effect of the factor (drug) that we are trying to study.

Drug (A or B) ----------→ cholesterol level ----------→ Heart attack or stroke

Similarly, if we wanted to know the effect of cholesterol on heart attack or stroke, we would not control for type of Drug. If cholesterol level is of interest, Drug is an exogeneous risk factor, not a confounder.

Collider - Artifactual relationships can appear even though there is no actual causation or association.

influenza infection ⟶ fever ⟵ food poisoning

Table 1 Data illustrative of selection bias, due to conditioning on a collider

	Influenza				
	Yes	No	Total	Risk	Risk difference
Panel A					
Sandwich					
Chicken	5	45	50	0.1	0.0
Egg salad	5	45	50	0.1	
Panel B					
Fever					
Sandwich					
Chicken	5	0	5	1.0	0.9
Egg salad	5	45	50	0.1	
No fever					
Sandwich					
Chicken	0	45	45	0.0	NA
Egg salad	0	0	0	NA	

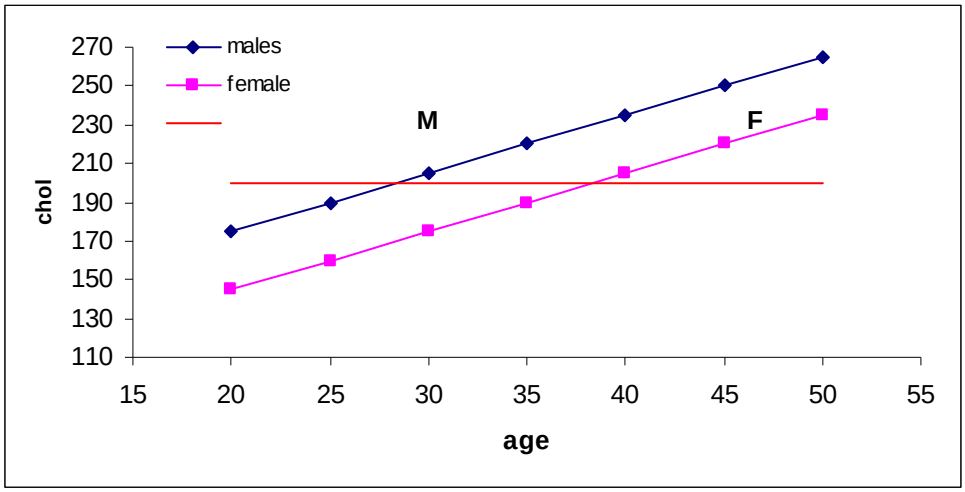
In the above example, there appears to be an association between contracting influenza and food poisoning since both are a cause of fever.

Confounding can both hide true differences and create artifactual differences

Ex 1- Cholesterol (mg/dl) in males and females - No apparent gender difference

variable	Males	Females
Mean age	30	40
Mean cholesterol	205	205

The mean cholesterol is the same in males and females but age is ignored.

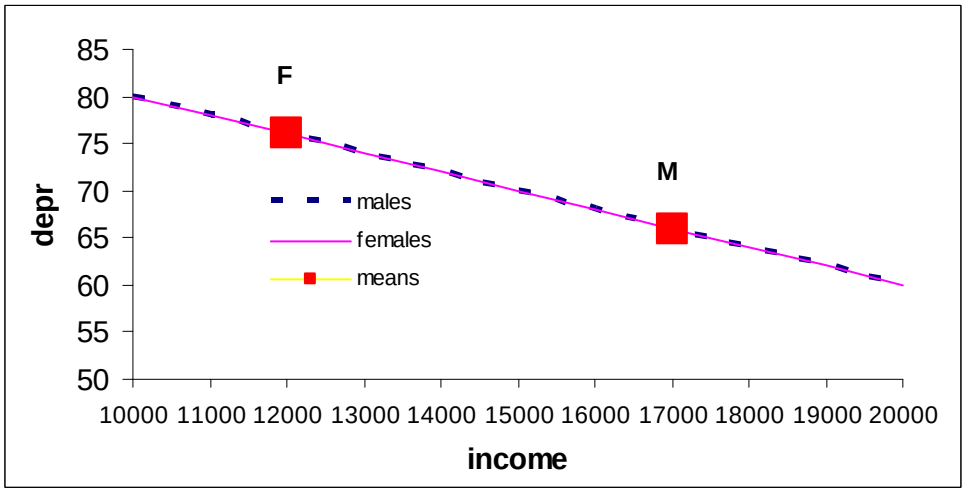


Controlling for age, the cholesterol means are not the same, males are higher than females

Ex 2 – Depression score in males versus females

variable	Male	Female
income	17000	12000
mean depr	66	76

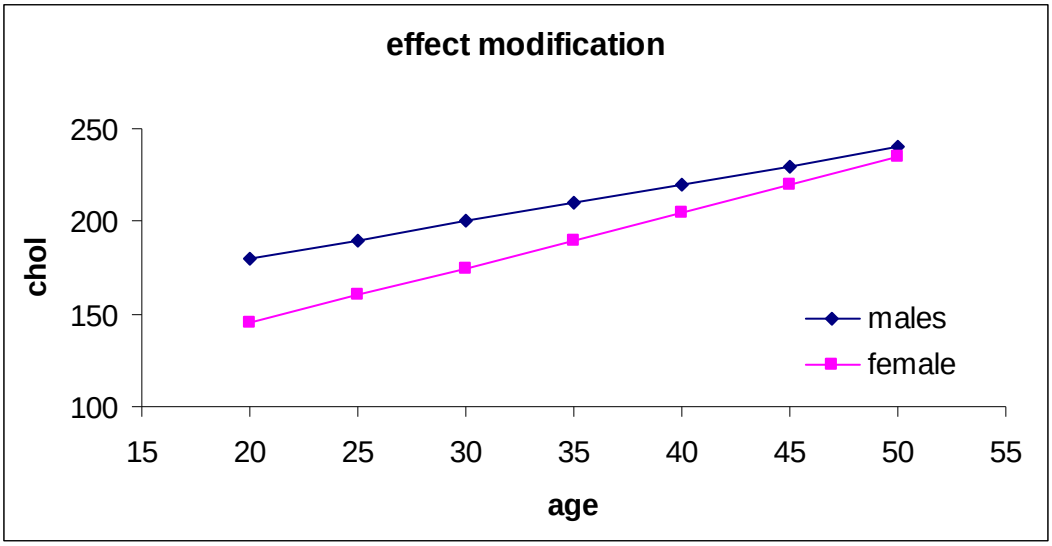
Males seem to have lower depression than females



Controlling for income, depression is the same in males and females

Effect modification (moderator, interaction)

When the effect of the risk factor is not the same for all levels of the confounder, the confounder is sometimes referred to as an effect modifier or a "moderator". This is also called a non parallelism, non additivity or an "interaction".



In the above (idealized) example, in younger individuals, cholesterol is higher on average in males compared to females. However, as persons age, the cholesterol gap narrows. By age 50, cholesterol is the same in males and females. Therefore, when one asks if cholesterol is different in males and females or if gender affects cholesterol, one must specify the age. The answer is not the same at all ages. The rate (slope) of increase in cholesterol per year is also not the same in males and females.
In psychology, age in the above example is a **"moderator"** of the effect of gender on cholesterol.

Relationships are not necessarily linear or additive

When investigating the effect of many factors, including potential confounders, on an outcome (Y), some investigators assume it is acceptable to look at each factor one at a time. This approach might only be minimally acceptable if the effects of all the factors on the outcome are linear and additive, although even then looking at factors one at a time is not optimal.

Many people assume (with no evidence) a "linear model"

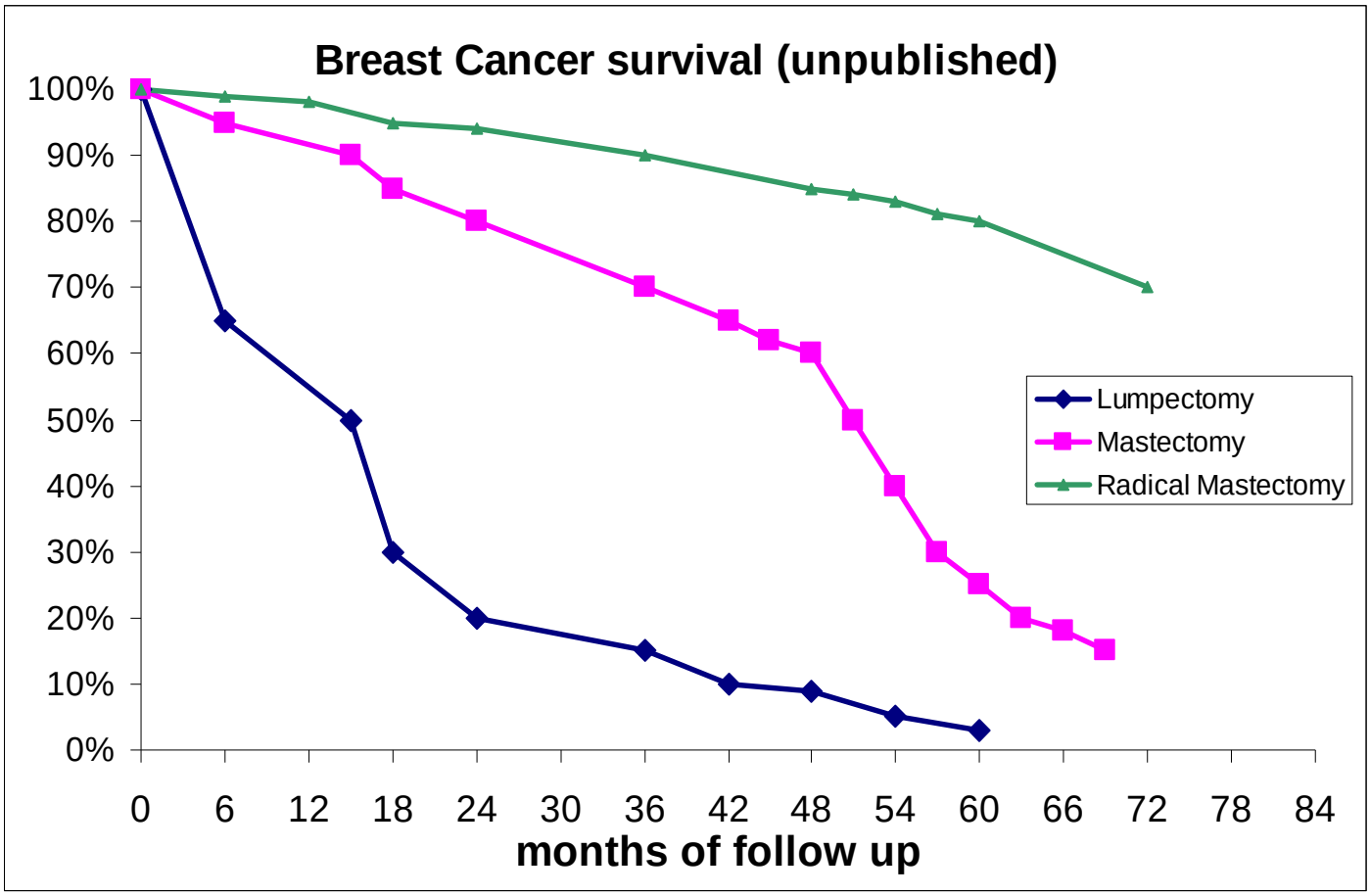
$$\text{Outcome(Y)} = b_0 + b_1 \text{ age} + b_2 \text{ gender} + b_3 \text{ SBP} + \ldots.$$

(ex: HDL = 46 + 0.15 age – 10 male)

But there is not reason to assume that relationships have this additive form. There are interactions

(symergisms, antagonisms).

Confounding - Is type of surgery the only reason these survival curves are different



Fisher et. al. Oct 2002 NEJM p1233

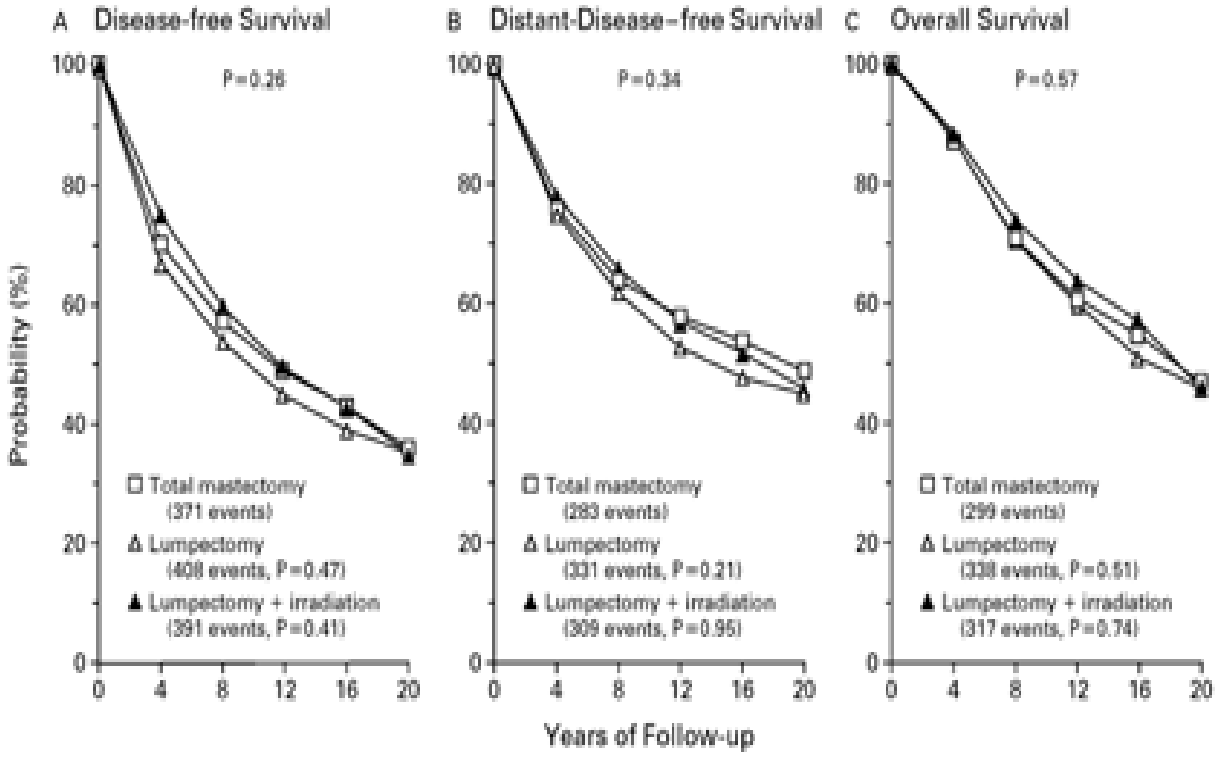
TWENTY-YEAR FOLLOW-UP OF A RANDOMIZED TRIAL COMPARING TOTAL MASTECTOMY, LUMPECTOMY, AND LUMPECTOMY PLUS IRRADIATION FOR THE TREATMENT OF INVASIVE BREAST CANCER

Background

In 1976, we initiated a randomized trial to determine whether lumpectomy with or without radiation therapy was as effective as total mastectomy for the treatment of invasive breast cancer.

Methods

A total of 1851 women for whom followup data were available and nodal status was known underwent randomly assigned treatment consisting of total mastectomy, lumpectomy alone, or lumpectomy
and breast irradiation. Kaplan–Meier and cumulative- incidence estimates of the outcome were obtained.



Reasons for lack of comparability – Bias

Unlike confounding, which is a factor present in the patients, bias is an often influence, factor or process **taken by the investigator (or ignored by the investigator)** which acts to make the observed results non-representative of the true effects of therapy or the true cause of a disease. Biases may mislead us when we extrapolate results to the general causes of disease or the treatment of future patients.

Some types of study design bias (**Ref- Boyce & Wilcox - Biomedical Bestiary**)

Selection bias (Berkson's bias)- The way subjects are selected for entry in the study differs in the various comparison groups and/or from the general population. Example: Season versus rate of complications for women in labor. Complication rate is based on hospital deliveries only. In a cold winter, only women with more difficult pregnancies might come to the hospital. While the true complication rate may be about constant, the self selection may cause an apparent increase in the complication rate in winter versus summer. This is an example of an "external" or "sampling" bias.

Variable observer bias - The apparent effect is due to a difference in the observers (ie. the MD) and not to a true difference in the outcome. Sometimes called “calibration” bias.

Hawthorne effect - The subject (patient) changes his response in the presence of the questioner (physician). Often just showing interest in a patient changes their response.

Diagnostic accuracy bias - The accuracy of the diagnosis changes (usually improves) over time. This can cause apparent disease incidence to change.

Response bias - The way and conditions under which the question is asked affect the answer. The Hawthorne effect is a specific response bias.

Lead time bias – Survival seems longer when disease is diagnosed earlier (screening)

Survival bias - Only those healthy enough to survive until data is collected can provide data.

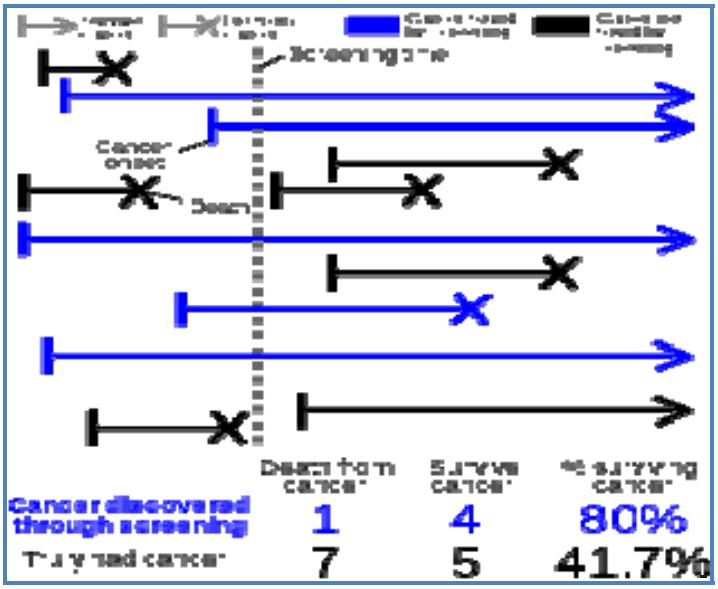
Example: Two hundred bladder cancer patients were randomized to two chemotherapeutic agents, A and B, with 100 patients assigned to each. Both were thought to have roughly equal efficacy, but the investigators wished to know if one was less toxic than the other. The white blood cell (WBC) count was used as a measure of toxicity. Treatments that are more toxic lower the WBC. The results are given below.

	Treatment A	Treatment B
Mean WBC	5600	4200
Sample size(n)	67	89

At first glance, it might seem that Treatment A has less toxicity. However, while 100 patients were assigned to each treatment, data was collected on only 67 patients on Treatment A and only 89 on Treatment B. Data could not be collected on patients who died. Therefore, it is no longer clear that A is better. A may in fact be associated with higher mortality. Only those alive can be measured.

Dropouts in a clinical trial are a major potential source of bias even though patients may be randomized to treatment. Should compare baseline characteristics in dropout vs non dropouts to see if dropping out is at random.

Length time bias – evaluating screening and early treatment



Those in a screening program may have their disease (cancer) detected earlier. This time is added to their "survival" time (time from diagnoses to endpoint) making it seem longer due to early treatment, not just earlier detection

NOTE: Bias is the same as a lack of **accuracy**. But a lack of accuracy is not the same as a lack of precision.

Some sources of bias:

Study design: Absence of a control group
Wrong type of controls used
Lack of control for other prognostic factors

Sample selection: Poor eligibility (inclusion/exclusion) criteria
Can't generalize to population of interest from "grab" (convenience) samples
Refusals – sickest persons may not agree to participate

Conduct of study: Differential **dropouts** – More/sicker dropouts in one group (like survival bias)
Poor and differential diagnosis and supportive care
Patients in treatment group get more attention than controls
Inadequate evaluation methods
Poor data quality, errors and missing data

Confounding versus bias

Unfortunately, in much of the medical literature, the terms confounding and bias are used interchangeably. However, this is usually not critical since the presence of either weakens the conclusions made.

Confounding and bias can occur in any kind of study. However, some study designs can reduce or eliminate confounding and bias.

Characteristics such as age, sex and the degree and length of illness are usually potential confounders. For example, in comparing treatments A and B, if the outcome is death, age is a confounder if the average age is not the same in groups A and B and if growing older increases your chances of dying independent of the choice of treatment.

The term **bias** is generally reserved for actions taken by the investigator in evaluating the value of the outcome. An outcome value is said to be biased when there is a systematic misclassification or mismeasurement of an experimental unit.

External bias/lack of validity (non representative sample)

The term "bias" is also used when the study sample is not representative of the target population (the population of interest). This is "external" bias or "selection" bias as noted above. Often, groups may be comparable within a study but results cannot be generalized to a wider population.

How to control for confounding

There are three main strategies for controlling confounding. The best is by using a good study design such as a randomized trial and by exclusion and inclusion criteria. The next best is by stratification and/or matching on confounders. The most complicated is by using statistical models such as regression models or propensity scores.

Regression example - neo adjuvant chemo vs cancer recurrence rate

In those successfully treated for breast cancer, when evaluating the association between giving neo-adjuvant chemotherapy versus breast cancer recurrence, the hazard rate ratio (HR) a statistic that indicates the association, is HR = 0.55 (p value=0.308), ignoring all possible confounders. That is, the recurrence rate is about half in those who had neo adjuvant chemotherapy. But, as shown in the table below, controlling for pre menopausal status (pre vs post), positive nodes (yes or no), cancer stage (1-4) and positive tumor margin (yes or no), the HR for neo adjuvant chemotherapy is now HR=0.20 (p value=0.048).

Regression model for breast cancer recurrence rate (Chang)

Predictor	HR	Lower conf bound	Upper conf bound	p value
Neoadj chemo	**0.20**	**0.04**	**1.00**	**0.0483**
Pre menopause	**2.12**	**0.72**	**6.25**	**0.1853**
Positive nodes	**3.64**	**1.29**	**10.28**	**0.0152**
Stage 2 vs 1	**8.30**	**2.11**	**32.59**	**0.0007**
Stage 3 vs 1	**2.67**	**0.42**	**17.07**	**0.3063**
Stage 4 vs 1	**46.83**	**7.94**	**276.3**	**< 0.0001**
Pos tumor margin	**3.14**	**1.15**	**8.59**	**0.0350**

Study designs and their effects on confounding and bias

Study designs - The choice of subjects and treatment assignments

A number of designs are used to try to control bias and confounding. They also have a number of other advantages and disadvantages listed below. Note, however, that not all of these designs are necessarily successful in eliminating bias or confounding. Moreover, in a given situation, not all of the designs are practical or ethical, even if they would be theoretically optimal. So what may be the "best" design scientifically may not be possible to actually carry out.

Experiments versus observational studies - Definition of an experiment

Experiments or "clinical trials" (an experiment carried out on human beings) are one type of study whose goal is to evaluate a treatment. The treatment does not have to be a drug, but may be a technique, surgical procedure or other type of therapy **.**

In a study where a number of treatments are being compared, the study is best classified as an **experiment or quasi experiment** if there was a **premeditated treatment assignment plan or treatment intervention or manipulation** on the part of the investigators and the **primary purpose** of this assignment plan is to evaluate the relative efficacy of the various treatments. That is, a study is classified as an experiment when the **main reason** for treatment assignment is to **make comparisons possible** and at least one of the treatments is not part of the standard therapy. We tend to use the term “quasi” experiment when there is a deliberate intervention under study but no randomization.

When treatments are not being evaluated or were applied for other reasons (such as a person was ill and would have received treatment anyway) the study is best classified as an **observational study**, not a planned experiment. Almost all epidemiology studies, or studies that try to establish the cause (etiology) of a disease in humans are observational studies. It is not possible for an investigator to assign disease risk factors, the investigator merely observes who has which factors.

Randomization can control confounding in experiments

In general, the best method of treatment assignment in an experiment is a **randomized** assignment. This is a method where every participant has an equal chance of being assigned to any of the treatment groups. This methods not only tends to make groups comparable with respect to known confounding factors but also makes them comparable with respect to **unknown** confounding factors. However, it is also important to realize that a randomized assignment to treatment is not always the most ethical method and should only be carried out when there is a true state of ignorance or a widespread controversy over which treatment is most effective and/or least toxic.

Blinding can control (internal) bias in experiments

Internal (non selection) bias can often be reduced or controlled by blinding (or masking) those providing the outcome information as to which treatment is being used. A study is called double blind or double masked if neither the patient nor the investigator making the treatment evaluation knows which treatment the patient is being given. Drug studies that are not blinded are often referred to as "open label" studies since the patient can "read the label" on the bottle and know what treatment she is getting.

Classification of studies and study designs -outline & examples

I. Experiments or clinical trials – Experiments usually address the question – does this novel treatment work or does this novel treatment or intervention have acceptable side effects? One can't use experiments to investigate causes of disease in humans**. Experiments are characterized by the investigator having control over who gets which treatment.** That is, treatment assignment is determined by the investigator's need to answer a scientific question about a novel treatment. In this strict definition, just using a standard treatment alone is not an experiment.

A. Randomized, controlled trial (RCT) – Usually the best design, patients are assigned at random to each treatment. Example: Breast cancer patients are randomized to surgery alone versus surgery plus chemotherapy.

B. Quasi Experiment or Parallel groups trial - A trial with independent concurrent controls but **no** randomization to treatment. Probably the most common type of clinical experiment. Example: The incidence of heart attacks is compared in persons taking a daily aspirin versus those not taking aspirin. Patients may decide whether to take aspirin or not.

C. Self controlled, before/After trial, paired trial – Same persons compared before versus after therapy or on two different treatments**.** Examples: The oral bacteria count is measured before and after use of a new mouthwash. Example: Two different facial acne treatments are given to teenagers, one on the right side and one on the left. The generalization of this is a **repeated measures** design. This design is also sometimes called a "case series" since there are no controls.

D. Crossover trial - One half of a group of patients are randomized to a treatment A while the other half are randomized to B. After a period of treatment (period 1) and a washout period, those initially receiving A are switched to B and those on B are switched to A. A crossover design can only be used for a **chronic** disease. When it is usable, it yields some of the most definitive evidence as it

combines the parallel group and the self control design.

E. Trials with Historical or External Controls - Example: Cancer survival in those treated with surgery and Herceptin after 1997 is compared to survival in those treated before 1997 with surgery alone, before Herceptin was created. Can't be sure that any observed differences are only due to the Herceptin as opposed to other factors that change over time.

F. Diagnostic assessment study – Diagnostic method A and diagnostic method B are compared relative to a "gold standard" where all patients are assessed by A, B and the gold standard. In this study, sensitivity and specificity are usually reported.

Cases studies are another type of design not included here.

Further examples:

Experiment – Randomized Clinical Trial (RCT)

Objective and Methods: A double-blind, randomized trial was conducted in 220 patients to assess the role of medical treatment for bleeding peptic ulcers. Patients with duodenal, gastric or stomal ulcers and signs of recent bleeding (confirmed by endoscopy) were randomly assigned to receive omeprazole (40 mg given orally every 12 hours for five days) or placebo. The outcome measures studied were further bleeding, surgery and death.

Results: Twelve of the 110 patients treated with omeprazole (10.9%) had continued bleeding or further bleeding, as compared with 40 of the 110 patients who received placebo (36.4%).
(Ref: N Engl J Med 1997: 336:1054-1058)

Quasi-Experiment

Objective and Methods: A prospective intervention study was conducted to assess the effect of a community-based program on changes in fracture rates and short-term hospital costs among people 65 years old or older. Residents of one municipality served as intervention subjects and residents of another municipality served as control subjects. Three years of baseline data were collected prior to the intervention, which consisted of removing environmental hazards from the home, promoting the use of safe footwear outdoors in winter etc. Add data were collected as part of a national injury surveillance system in Norway.

Results: The overall fracture rate fell from 34.98 per 1000 person-years to 31.58 per 1000 person-years in the intervention municipality. The rate increased from 27.83 per 1000 person-years to 36.92 per 1000 person-years in the control municipality. In the intervention municipality, the short-term hospital admission rate fell from 8.99 per 1000 person-years during baseline to 7.54 per 1000 person-years during the intervention period. (Ref: J Epidemoil Community Health 1996; 50:551-558)

Advantages and disadvantages of Experiments / clinical trials (See Hully & Cummings)

Disadvantages

Experiments are very costly in time and money.

Many research questions can't be addressed because of ethical problems or the disease is too rare

Physicians and patients often unwilling to participate, particularly in randomized trials.

Inappropriate use of historical controls or no controls can produce major errors! (less of a problem with concurrent controls)

Answers from standardized clinical trials may be different from the behavior in general practice. For example only a single fixed dose may be evaluated in a trial, whereas the general practice uses many doses.

Trials tend to restrict the scope and the questions under study.

Advantages

Properly controlled and designed Experiments produce strongest evidence for cause & effect or lack thereof. May be unethical to give a treatment that does not work. Important in an era of proliferating medical technology.

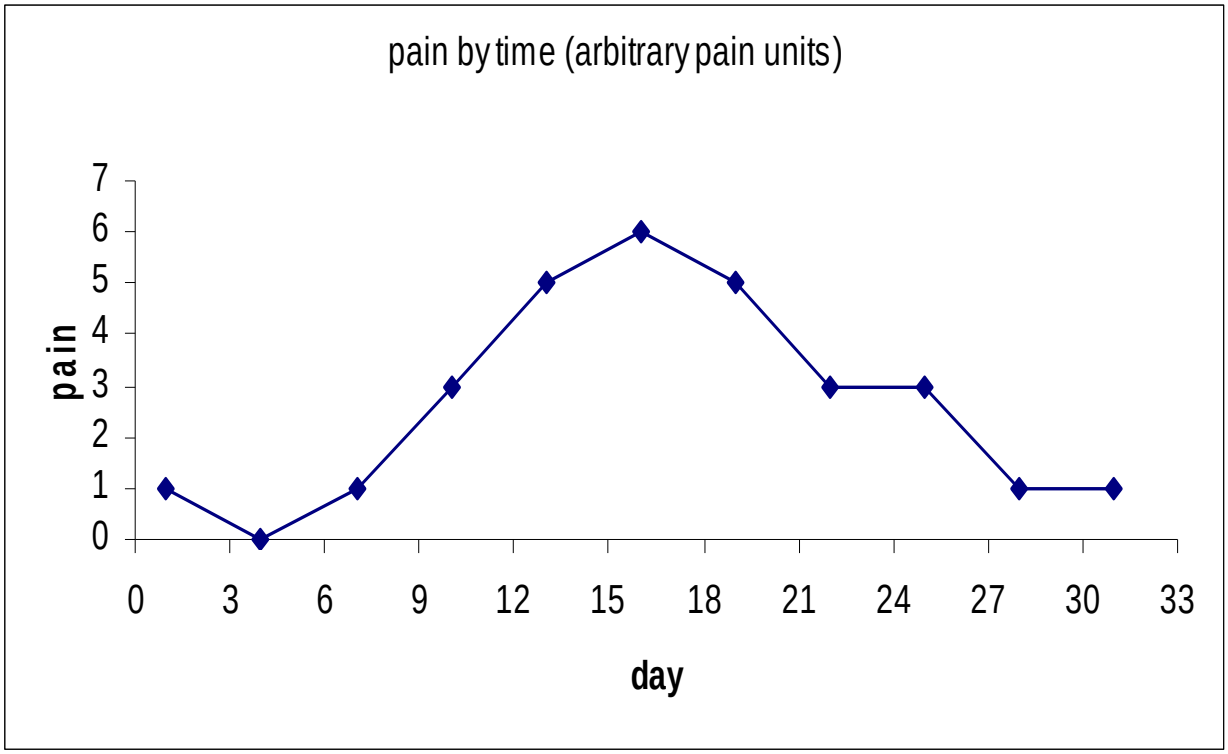
Randomized trials are best for assuring **comparability** and best for controlling confounding and bias.

Sometimes required by the Govt. (FDA and new drugs)

Can be faster and cheaper in the long run if they put a controversy to rest.

Self controlled Before and after paired design – no control group

Example – non conventional treatment for pain.



Since patients seek treatment when their pain is worse, apparent pain reduction is not due to the treatment but is simply the natural history of the disease. Pain "comes and goes" anyway.

The Factorial Design (Factorial Experiment)

Synergism, Antagonism, Additivity

When one is investigating several treatments simultaneously, an experiment where all possible combinations of treatments are given with one combination assigned per group is called a **factorial design or factorial experiment**. This is often done with animals.

For example, if we are investigating three treatments, A, B and C, where A has three levels (for example, low, medium, high), and B and C each have two levels (for example, given or not given) a full factorial experiment would have 3 x 2 x 2 = 12 groups. (Also called design "cells").

No C

	No B	B
Low A	Outcome	Outcome
Medium A	Outcome	Outcome
High A	Outcome	Outcome

C

	No B	B
Low A	Outcome	Outcome
Medium A	Outcome	Outcome
High A	Outcome	Outcome

If there are the same number of subjects in each of the groups, this is termed a **balanced** factorial experiment. Otherwise it is unbalanced. Balanced experiments have certain convenient statistical properties but are not necessarily required. However, if there is a randomized assignment to the treatments, the sample sizes should be (at least approximately) balanced.

Example 1 (above): In immunotherapy for lung cancer, low, medium or high doses of cytotoxic T cells are given either with or without Interleukin-2 in male or female adults. The outcome is the reduction in tumor volume as shown by changes in the radiograph post minus pre treatment. There is a separate randomization schedule for males and females. The advantage of the factorial design is that **interactions** of the factors can be evaluated.

Example 2: Two treatments (anti arrhythmic drug, high dose NSAID) in addition to the standard treatment are given to persons who have had a recent MI (2 x 2 design).

Survival at 3 years in MI patients on standard treatment plus anti arrhythmic and/or NSAID

treatment	low dose NSAID	high dose NSAID
no anti arrhythmic tx	**60%**	**80%**
anti arrhythmic tx	**70%**	**65%** (might expect 90% if additive)

The interaction here is an antagonism. The treatment combination does not increase survival as much as would be expected by the additive combination of each treatment. (The opposite is a synergism). In this example both treatments together are worse than either alone.

The repeated measure design

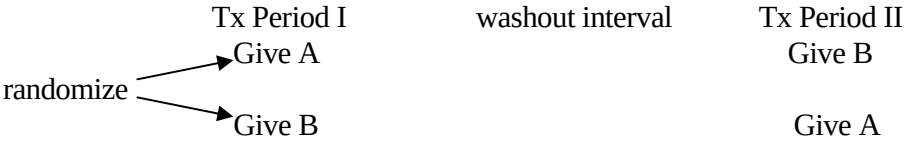
Each subject measured repeatedly over time. A **paired** comparison is a special case of this design where there are only two times. In the example below, treatment is the "between group" factor, and time is the "within group" factor since the same subject is measured four times. Measuring the same subjectfour times is is NOT the same as measuring four different groups of subjects. The between group and within group comparisons have different statistical properties.

	Time 1	Time 2	Time 3	Time 4
Treatment A	Y	Y	Y	Y
Treatment B	Y	Y	Y	Y

Cross over design and its pitfalls

One of the best experimental designs for comparing two or more treatment is the so called "cross over" design, so called because each patient "crosses over" to the other treatment(s). In the two period cross over, half of the participants are initially randomized to treatment A and the other half are initially randomized to treatment B.

Two period cross over design



In a typical analysis, once the trial is concluded, one checks to see that the timing (Period) and order of administration (AB or BA) does not affect the results. That is, the effect of A relative to B should not depend on order. Moreover, the effect of A (or B) at period I should be the same as that of A (or B) at period II. If the order of administration or the time the treatment was given has an influence on the outcome, one says that there are nuisance carry over effects or period effects respectively.

Example: Timolol versus placebo

Consider a clinical trial for the drug Timolol versus placebo for the treatment of chronic migraine headaches. (Note that a cross over trial can only be conducted on persons with chronic diseases). Within a treatment period, a patient either experiences complete relief (headache gone, yes or no) or not. If there is no period or carry over effects, the results may be as below.

Percent who obtain relief by order of treatment – no period or carry over effects, only a treatment effect

Treatment order	Period 1	Period 2
AB	43% (Timolol)	27% (placebo)
BA	27% (placebo)	43% (Timolol)

Notice that the difference in relief for Timolol versus placebo is 43%-27% = 16% in this example regardless of order of administration.

If there is a **period effect** but no carry over effects, the results could be as below.

Percent who obtain relief by order of treatment – treatment and period effects

Treatment order	Period 1	Period 2
AB	43% (Timolol)	37% (placebo)
BA	27% (placebo)	53% (Timolol)

In this example, unknown to the investigator, the treatment improves relief by 16% but just going from period 1 to period 2 also increases relief by an additional 10%.

If one looks at the AB group only, one sees only an apparent 43%-37%= 6% improvement due to Timolol. If one looks at the BA group only, one sees an apparent 53%-27%= 26% improvement. The true improvement of 16% is the average of 6% and 26%. ([26%+6%]/2= 16%). Note also, that the between group differences still show a 16% improvement due to Timolol in either Period. So, while having a period effect may be a bit confusing, it is still possible to get unbiased (but less efficient) estimates of the true drug effect, provided one recognizes the simultaneous influence of a period effect.

Finally, the most serious problem is a **carryover (or order) effect** as illustrated below.

Percent who obtain relief by order of treatment – treatment and carryover (order) effects

Treatment order	Period 1	Period 2
AB	43% (Timolol)	41% (placebo)
BA	27% (placebo)	43% (Timolol)

Notice that, if we compare both groups using the Period 1 data only, we see a 16% improvement due to Timolol. But, if we look only at those who got Timolol first (the AB group only) the apparent effect of Timolol is only 43% - 41% = 2%. This is because, unknown to the investigator, Timolol made a permanent change (a "cure") in some of those treated. As before, in the BA group, the apparent effect of Timolol is 16%. However, now the average of 2% and 16% does not give the true effect of 16%, but an apparent effect of 10%. Note that here, unlike the period effect case, the difference between the two groups at each period is no longer 16%. Rather, the difference is 16% at Period 1 and 2% at Period 2.

So, in the case of a carryover or order effect, only the data at Period 1 gives an unbiased estimate of the true effect.

Criteria for the “best” experiments/trials

(Bausell R, Snake Oil Science, Oxford Univ Press,

Bausell suggests that there are at least five criteria that make a study truly outstanding and very likely to provide a true and reliable answer to whether a treatment works.

1. Randomized Trial

2. Double blind (if applicable)

3. Large sample size (at least 50/group)

4. No more than 25% dropouts in any group

5. Published in a high quality peer reviewed Journal

He would say that such studies provide the best evidence and should be preferred over results from studies which do no meet these criteria when results do not agree.

Observational studies

II. Observational studies – Usually address the question what (exposure) is the cause of the disease? (Epidemiology) In observational studies, the investigator has **no** control over treatment or exposure to disease. It is generally not possible or ethical to perform experiments to determine causes of disease.

A. Cohort, prospective or longitudinal study - a "natural" experiment. (Often similar to a quasi – experiment). Ascertainment is by exposure and is longitudinal. Example: Groups of Gulf War veterans are followed from the beginning of their military service to see if those who were (inadvertently) exposed to fumes from burning oil fields have the same incidence of cancer as those not exposed.

A1 Historical Cohort study (not seen often - a variant of a cohort study that is often confused with a case-control study since it has a "retrospective" quality")

Example: Blood is routinely collected on all women giving birth in a certain hospital. The blood is thawed later and DDT levels are obtained and related to the incidence of cancer.

B. Cross sectional (A survey or "medical poll") Ascertainment is cross sectional. Example: Menopausal women from a church group are surveyed as to whether they are using estrogens during menopause and are questioned regarding the degree of menopausal discomfort they are experiencing.

C. Retrospective or case-control (the most common epidemiological design) Ascertainment is by disease (yes or no). Example: One group of men with AIDS and another without are asked about their past drug use.

D. **Ecologic Studies** - A larger aggregate (not person) is the unit of analysis – Generally cross sectional.

Some study designs are better able to establish causality than others. For example, a retrospective study examining a possible link between depression and cancer might compare rates of depression in those with and without cancer. But this study cannot necessarily determine if the depression caused the cancer or the cancer caused the depression. Only a prospective study can rule out the latter possibility.

In general, case-control studies give no information on temporal ordering.

Often, an observational study (particularly an inexpensive case-control study) is done first to generate hypotheses and gather "pilot" data for planning purposes. Then a longer, more expensive and hopefully more "definitive" study (usually a cohort study or clinical trial) may be undertaken.

Question - Why would one not want to do a case-control study regarding the relation between depression and cancer?

Observational studies – Examples that contradict each other

Cohort

Coffee and alcohol consumption and the risk of pancreatic cancer in two prospective United States cohorts. **Michaud DS**, **Giovannucci E**, **Willett WC**, **Colditz GA**, **Fuchs CS**. Cancer Epidemiol Biomarkers Prev. 2001 May;10(5):429-37.

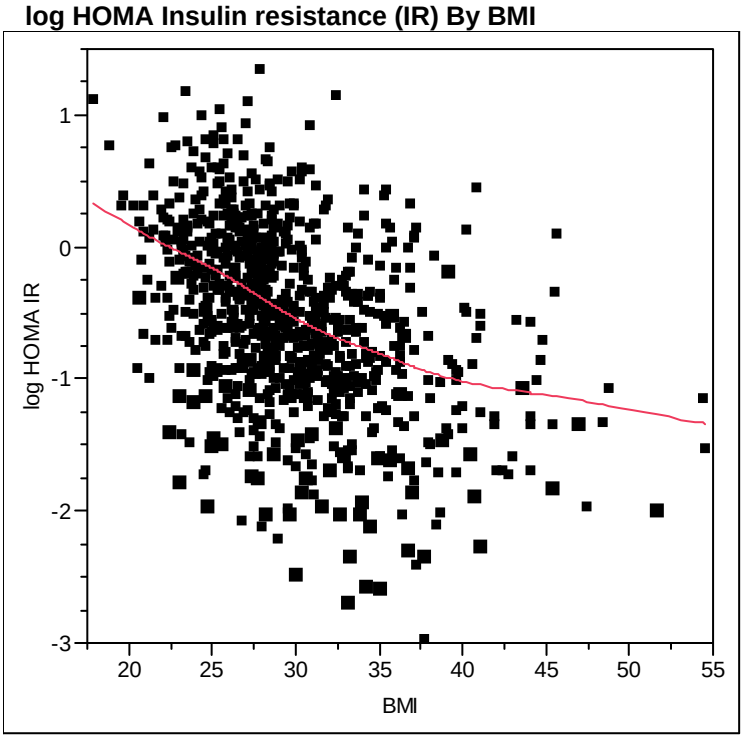
Although most prospective cohort studies do not support an association between coffee consumption and pancreatic cancer, the findings for alcohol are inconsistent. Recently, a large prospective cohort study of women reported statistically significant elevations in risk of pancreatic cancer for both coffee and alcoholic beverage consumption. We obtained data on coffee, alcohol, and other dietary factors using semiquantitative food frequency questionnaires administered at baseline (1986 in the Health Professionals Follow-Up Study and 1980 in the Nurses' Health Study) and in subsequent follow-up questionnaires. Data on other risk factors for pancreatic cancer, including cigarette smoking, were also available. Individuals with a history of cancer at study initiation were excluded from all of the analyses. During the 1,907,222 person-years of follow-up, 288 incident cases of pancreatic cancer were diagnosed. The data were analyzed separately for each cohort, and results were pooled to compute overall relative risks (RR). Neither coffee nor alcohol intakes were associated with an increased risk of pancreatic cancer in either cohort or after pooling the results (pooled RR, **0.62**; 95% confidence interval, 0.27-1.43, for >3 cups of coffee/day versus none; and pooled RR, 1.00; 95% confidence interval, 0.57-1.76, for > or = 30 grams of alcohol/day versus none). The associations did not change with analyses examining different latency periods for coffee and alcohol. Similarly, no statistically significant associations were observed for intakes of tea, decaffeinated coffee, total caffeine, or alcoholic beverages. Data from these two large cohorts do not support any overall association between coffee intake or alcohol intake and risk of pancreatic cancer.
PMID: 11352851 [PubMed - indexed for MEDLINE]

Case-Control

Coffee and cancer of the pancreas **MacMahon B**, **Yen S**, **Trichopoulos D**, **Warren K**, **Nardi G**. N Engl J Med. 1981 Mar 12;304(11):630-3.
We questioned 369 patients with histologically proved cancer of the pancreas and 644 control patients about their use of tobacco, alcohol, tea, and coffee. There was a weak positive association between pancreatic cancer and cigarette smoking, but we found no association with use of cigars, pipe tobacco, alcoholic beverages, or tea. A strong association between coffee consumption and pancreatic cancer was evident in both sexes. The association was not affected by controlling for cigarette use. For the sexes combined, there was a significant dose-response relation (P approximately 0.001); after adjustment for cigarette smoking, the relative risk associated with drinking up to two cups of coffee per day was 1.8 (95% confidence limits, 1.0 to 3.0), and that with three or more cups per day was **2.7** (1.6 to 4.7). This association should be evaluated with other data; if it reflects a causal relation between coffee drinking and pancreatic cancer, coffee use might account for a substantial proportion of the cases of this disease in the United States. PMID: 7453739 [PubMed - indexed for MEDLINE]

Cross sectional study - (log) Insulin resistance versus body mass index (BMI)

MESA (Multi-Ethnic Study of Atherosclerosis**)** data – FY 2000 (baseline)

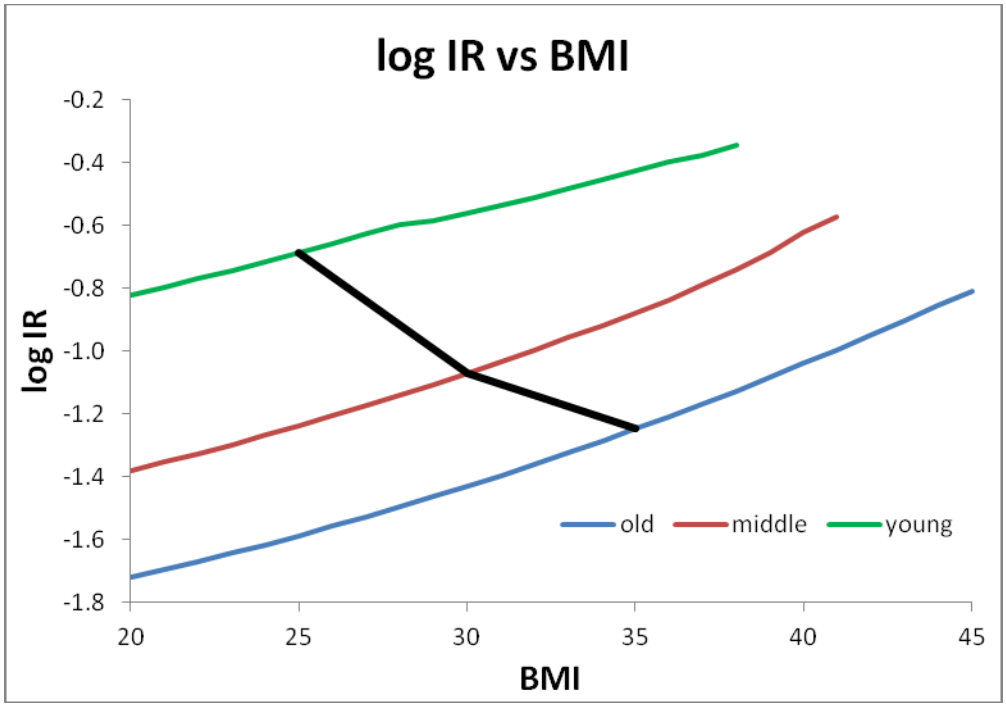


$n=750,\ r= -0.45,\ r_s= -0.46,\ p < 0.001$

The results above imply that (log) insulin resistance decreases as body mass index increases. However, this assumes that the cross sectional relationship seen above would be the same if individuals were followed up longitudinally. ("Young Cuban men in Miami grow up to be old Jewish men").

In a longitudinal (cohort) study, HOMA IR increases with BMI!

Cohort effect



The previous misleading results (trend line going down) from the cross sectional study can be due, at least in part, to what is called a "cohort" effect. In the above example, those born in a younger (later) cohort have higher risk of insuling resistance due to age and possibly other factors associated with the cohort in which they are born.

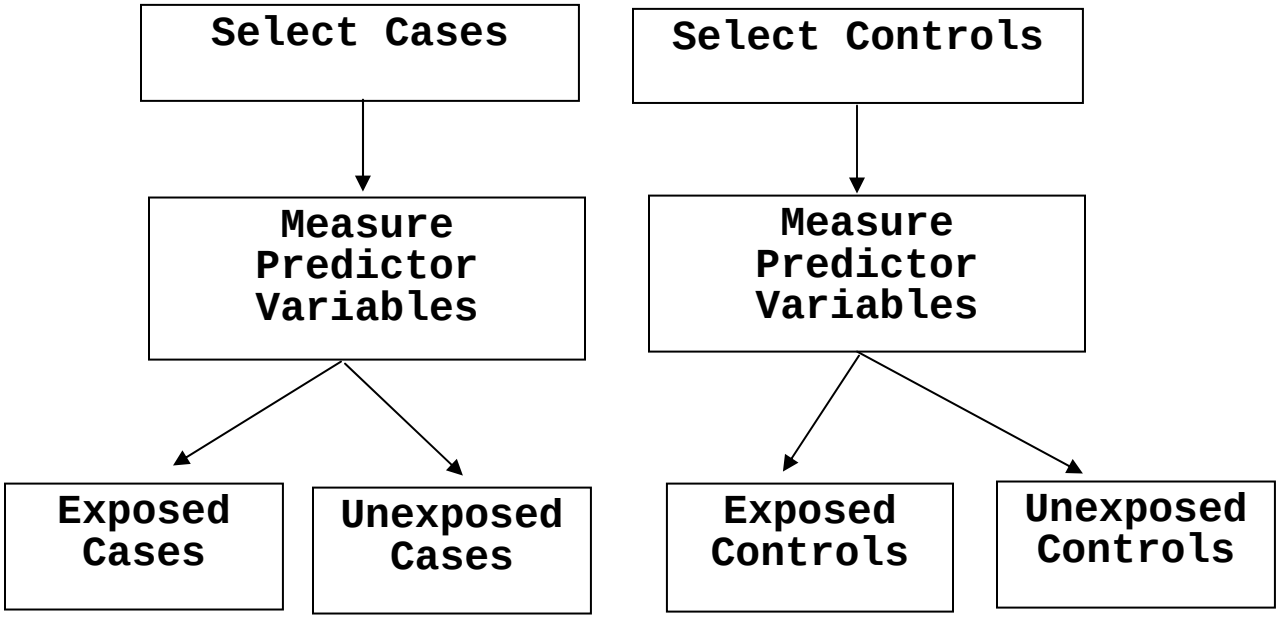
Controlling for this, the relation between BMI and insuling resistance is positive.

Example:

Case control study

Objective and Methods: The objective of the study was to assess the relationship between a T594M point mutation (the most common sodium-channel mutation) and hypertension among blacks. Cases were 206 black patients with high blood pressure (on treatment or with consistent supine systolic blood pressure of at least 140 mm Hg or diastolic blood pressure of a least 90 mm Hg) taken from referrals to a hypertension clinic in London between February 1995 and August 1996. The 142 normotensive controls were taken from a survey of the population aged 49-59 in the area from which the cases originated. All subjects were screened form T594M. PCR was used to amplify the epithelial sodium-channel β subunit from genomic DNA; the T594M variant was detected by single strand conformational polymorphism analysis of PCr products and confirmed by DNA sequencing.

Results: Of the 206 hypertensive cases, 17 (8.3%) had the T594M variant mutation compared to 3 of the 142 controls (2.1%). (Ref: Lancet 1998; 351:1388-92).



Advantages and disadvantages of observational studies

Main disadvantage for all observational studies – One can't assure comparability. Mostly used for epidemiology, for determining risk factors for disease. Main advantages for most observational studies- No ethical problems, generally lower expense per study.

Design	**advantages**	**disadvantages**
COHORT		
	Establishes sequence of events	Usually need large sample size
	Avoids bias in measuring predictors	Not feasible for rare outcomes/diseases
	Avoids survival bias	May have long duration
	Can study several outcomes	May have problems with dropouts/loss to follow up
	Yields incidence, relative risk, risk difference	
	Gives control of selection of subjects and over what to measure	Does not guarantee comparability
	Outcome status is not likely to affect the measurement of the exposure.	
	Outcome status is not likely to affect the selection of subjects (no selection bias)	
CROSS-SECTIONAL		
	May study several outcomes	Does not establish temporal order
	Can study several exposures	Exposure information collected from memories may not be accurate (recall bias)
	Gives control over selection of subjects and what to measure	Subject to survival bias
	Short duration	Not feasible for rare diseases
	Yields prevalence (not incidence)	Can't distinguish between predictors of disease occurrence vs disease progression
	Can be first step of a cohort study	**Assumes observed associations between persons are the same as associations within persons**
		Can't provide incidence
CASE-CONTROL		
	Useful and feasible for rare diseases	Possible bias from sampling two populations (ie getting appropriate controls may be difficult)
	Short duration	Does not establish temporal order
	Inexpensive, needs little labor	Often has bias when measuring potential predictors (example: recall bias)
	Can look at many risk factors at once	Subject to survival bias
		Can't estimate incidence or prevalence

The case control study is the easiest to do and the most vulnerable to bias and confounding.

Exploratory versus Confirmatory studies for experiments and observational studies

Another important distinction that applies to both experiments and observational studies is the distinction between an exploratory, or hypothesis generating study versus a confirmatory, or validation study.

The exploratory study does not have as rigid a set of hypotheses and has more general study questions. The criteria for claiming that an association or relationship has been found is generally more liberal. However, the results for this study are generally not considered proven until the corresponding validation study. (a "fishing expedition").

The confirmatory study has very specific hypothesis / models based on the previous exploratory study. There are generally much more strict criterion for proving "significant" relations in this study.

In drug trials, the Phase I and II trials are generally more exploratory and the Phase III trial(s) is (are) supposed to be confirmatory.

In observational studies, typically a statistical model, such as a regression equation model, is generated in the exploratory study after investigating several (perhaps many) models. But a subsequent validation study is needed to confirm this model and properly assess its accuracy.
Similarly, in exploratory etiologic studies, many possible causes of a disease may be investigated and only those that appear important are singled out for further study. In this sense, the exploratory study has a "screening" function.

Controlling for confounders (prognostic factors) - stratification

An important predictor of survival in leukemia is age at diagnosis. Persons who are younger at diagnosis are known to have better survival than older persons. Age by itself is termed a **prognostic factor.** However, when the focus is on evaluating a treatment, such as bone marrow transplant therapy, if age is not controlled for, age becomes a **confounder.** Below are listed three types of fallacies that can arise if the influence of the prognostic factor is ignored. (Taken from Dr Karim Hirji).

Assume that we are evaluating two treatments. Treatment A is a bone marrow transplant from oneself. Treatment B is a bone marrow transplant from a first degree relative. One treatment or the other is given after full body radiation for leukemia. The **outcome** is the survival rate (percent alive) after 3 months post radiation and transplant. In the examples below, the overall comparison of A versus B **ignoring** age is given first. Next a comparison stratified by age (young or old) is given.

I. False effect - treatment A appears to work better than B

Treatment	Alive	Dead	Total
A	74 (74%)	26	100
B	26 (26%)	74	100
	younger only		
A	72 (90%)	8	80
B	18 (90%)	2	20
	older only		
A	2 (10%)	18	20
B	8 (10%)	72	80

In this example, once age is considered, treatment A no longer appears to be better. Younger persons tended to get treatment A.

II. Treatment efficacy obscured (Simpson's "paradox")

Treatment	Alive	Dead	Total
A	50 (50%)	50	100
B	50 (50%)	50	100
	younger only		
A	30 (75%)	10	40
B	48 (60%)	32	80
	older only		
A	20 (33%)	40	60
B	2 (10%)	18	20

In this example, overall results imply A is equivalent to B. However, when stratified by age, A is superior to B in both strata.

III. Interaction

Treatment	Alive	Dead	Total
A	60 (60%)	40	100
B	60 (60%)	40	100
	younger only		
A	54 (90%)	6	60
B	36 (60%)	24	60
	older only		
A	6 (15%)	34	40
B	24 (60%)	16	40

Again, overall the two treatments do not appear different. However, treatment A seems to be better for younger patients and treatment B seems to be better for older patients. Thus the treatment can have opposite effects.

These examples illustrate why one must either try to equalize the distribution of a prognostic (or potentially confounding) factor or control for it. Randomization or randomization within strata to treatment is one way by which this is done. Matching and cross-over designs are other ways.

Some statistical methods for controlling for confounding

Stratification

Rate adjustment

Regression analysis

Propensity scores

These method can be used when randomization, pairing or matching in the design cannot be done. Propensity scores are a type of matching.

Adjusting rates- Controlling for confounding

(an example of stratification and controlling confounding)

Friedman reviews a study of possible sex bias at UC Berkeley in 1973. Of the 2691 men who applied for graduate admission to the six most popular majors, 1198 or 44.5% were admitted. Of the 1835 women who applied to these same six majors, only 557 or 30.3% were admitted. Therefore, charges of sex bias were brought against the university.

At UC Berkeley, the decision on admission is made within each department. The table below shows the admission breakdown by major (i.e. by department). The majors are identified by letter only since confidentiality policy did not allow the names of the majors to be released.

Admission by major at UC Berkeley, Fall 1973

Major	men		women	
	number applied	pct admitted	number applied	pct admitted
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	373	6%	341	7%
overall	2691	44.5%	1835	30.3%

There is an apparent paradox here. While overall, a higher percentage of men were admitted compared to women, when broken down by major, there was a higher percentage of women admitted in majors A, B, D and F. Only departments C and E admitted a higher percentage of men and the difference is not too large. Only 3% more men were admitted in major C and only 4% more were admitted in major D. Overall, there was about a 14% difference in favor of men. How do we account for this aparent contradicrtion?

The key to resolving the apparent contradiction lies in noticing that **equal numbers of men and women did not apply to each major and the admission rate was not the same for each major.** Over 50% of the men applied to the first two majors, which were fairly easy to get into for either sex. But over 90% of the women applied to the last four majors, and they were harder to get into.

It therefore might be misleading to compare the "crude" overall admission rates between men and women. Instead, we might consider a "**weighted**" rate,(also called a weighted average) where we weigh each majors **specific** rate by the majors share of the total number of applicants. We can use the **same** weights in both mean and women. This allows us to compute the overall admission rates as if equal numbers of mean and women had applied to each major. The table below shows the total number who applied to each major and is used to compute the weighted rates.

Total number of applicants to each major

Major	total number of applicants	percent of total
A	933 (825 + 108)	20.6%
B	585 (560 + 25)	12.9%
C	918 (325 + 593)	20.3%
D	792 (417 + 375)	17.5%
E	584 (191 + 393)	12.9%
F	714 (373 + 341)	15.8%
total	4526	100.0%

The weighted average for the men is

$$\frac{933 \times 62\% + 585 \times 63\% + 918 \times 37\% + 792 \times 33\% + 584 \times 28\% + 714 \times 6\%}{4526} = 39\%$$

Using the same weights, the weighted average for the women is

$$\frac{933 \times 82\% + 585 \times 68\% + 918 \times 34\% + 792 \times 35\% + 584 \times 24\% + 714 \times 7\%}{4526} = 43\%$$

These adjusted rates (adjusting for differing admission to each major) show that **adjusted** admission rates are about the same for men and women.

In vital statistics, the same technique is used. We combine **age specific** rates into overall **age adjusted** rates. The role of age strata is the same as the role of major in this example. When no adjustment is done, the rates are called **crude** rates. In the above example, the 44.5% admission for men and 30.3% admission for women would be the crude admission rates.

Outline for assessing an article in the Biomedical literature (Adapted from Colton: Statistics in Medicine)

I. Objectives

a. What is the goal or purpose of the study? What scientific hypothesis is being tested?

b. What is the target population – to whom do the investigators wish to apply the results? Who was included and excluded?

II. Study design

a. Is the study a planned experiment, quasi experiment or observational study?

b. What is the population from which the sample was selected?

c. How was the sample selected/participants chosen? Are their sources of bias? Are reasons for inclusion and exclusion of study subjects defined?

d. If the study was an experiment, were the subjects randomly assigned to treatment? Was the randomization scheme stated?

e. Was there an adequate control group?

f. Are the groups comparable at baseline?

g. Was there a sample size / power calculation carried out as part of the study planning?

III. Observations

a. What are the outcome measures? Are they clearly defined?

b. What are the predictors and relevant covariates?

c. Are the measures reproducible (reliable) and understandable?

IV. Analysis

a. What statistical hypotheses are being tested? Is this consistent with the goals in part I?

b. What type of analyses and statistical tests were performed? Are the calculations correct? Are the analysis methods consistent with the nature of the data?

c. What assumptions have been made about the data or design? Are they reasonable?

d. Have important, relevant factors and extraneous influences been accounted for in the analysis? Were confounding factors controlled?

e. Were the analysis results properly interpreted?

f. Were negative results distinguished from inconclusive results? Was the sample size large enough?

V. Presentation

a. Are the data and findings presented clearly? Is there sufficient detail to allow the reader to judge them?

b. Are the findings internally consistent? Do numbers add up and match in various tables and figures?

VI. Conclusions

a. What conclusions do the investigators draw? Do they exceed the data presented?

b. Do the conclusions related to the goals of the study? Do they answer the study questions?

VII. Redesign/ reanalysis

If parts of the design or analysis are thought to be inadequate, how would you would redesign the study and/or reanalyze the data. Be practical. That is, recognize that there are financial, time and ethical limits

to the types of studies that can be carried out.

Take home ideas

When trying to establish an association we usually ask at least two questions:

1. If several groups are being compared, are they comparable for all factors but the ones under study? In an experiment, is the treatment the only relevant factor that differs?

2. If there is comparability, is it likely that the association we observe is due to random variation only? Small differences may not be due to any true association, but only to random subject-to-subject variation and/or measurement error.

In addition, we often ask two additional questions.

3. If we observe an association from a sample, is it generalizable? That is, is our sample similar to the larger universe of population from which is was obtained? (Can we even identify the universe to which we might generalize?).

4. If we fail to find an association in a study, can we conclude that the association does not exist? Or is the association just to weak to detect? (Perhaps a larger sample would have greater sensitivity or power, and allow us to detect weaker associations).

Association is necessary but not sufficient to establish causality. To better establish causality, (A causes B) one must have correct temporal and/or dose order and be able to assess effect of A given that all other possible effects on B have been controlled for. (All else equal). Different study designs do a better or worse job in controlling for all other factors.

To establish an association, one must show that it is larger than would be expected by random variation. Moreover, one must show that there was sufficient sensitivity in order to show the absence of an association.