Statistical Literacy for Medical Librarians: Swimming in a Whirlpool of Conflicting Medical Claims

Steve Simon, P.Mean Consulting

Sign up for a free newsletter about Statistics, The Monthly Mean.
www.pmean.com/news

2. Outline of class

1. Show me your proof: Confidence intervals and p-values.
2. How bad is it really? Measures of risk.
3. Putting your life in the hands of a coin: Randomized trials.
4. It’s just what the doctor ordered: observational studies.
5. Putting it all together: Meta-analyses and systematic overviews.

We will have breaks mid-morning, lunch, and mid-afternoon.

3. My teaching philosophy

- No formulas.
- You are a consumer of Statistics.
- You’re smarter than you think.
- Ask questions at anytime.
- We will have scheduled breaks, but you can get up at anytime for refreshments or just to stretch.

4. Show me your proof: Confidence intervals and p-values

- Abstract: P-values and confidence intervals are the fundamental tools used in most inferential data analyses. They are possibly the most commonly reported statistics in the medical literature. Unfortunately, both p-values and confidence intervals are subject to frequent misinterpretations.

5. Show me your proof: Confidence intervals and p-values

- Abstract (cont'd): In this presentation, you will learn the proper interpretation of p-values and confidence intervals, and the common abuses and misconceptions about these statistics. You will also see a simple application of Bayesian analysis which provides an alternative to p-values and confidence intervals.

6. Learning objectives

- In this section, you will learn how to:
  - distinguish between statistical significance and clinical significance;
  - define and interpret p-values;
  - explain the ethical issues associated with inadequate sample sizes.

7. Outline

1. Pop quiz
2. Definitions
3. What is a p-value?
4. Practice exercises
5. What is a confidence interval?
6. Practice exercises
7. Repeat of pop quiz

8. Pop quiz #1

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

9. Pop quiz #2

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

1. The result is statistically significant and clinically important.
2. The result is not statistically significant, but is clinically important.
3. The result is statistically significant, but not clinically important.
4. The result is not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

10. Definitions: Population

- A population is a collection of items of interest in research. The population represents a group that you wish to generalize your research to. Populations are often **defined in terms of demography, geography, occupation, time, care requirements, diagnosis, or some combination of the above**.

11. Definitions: Population

- An example of a population would be all infants born in the state of Missouri during the 1995 calendar year who have one or more visits to the Emergency room during their first year of life.

12. Definitions: Sample

- A sample is a subset of a population. A random sample is a subset where every item in the population has the same probability of being in the sample. **Usually, the size of the sample is much less than the size of the population**. The primary goal of much research is to use information collected from a sample to try to characterize a certain population.

13. Definitions: Type I Error

- In your research, you specify a null hypothesis (typically labeled H0) and an alternative hypothesis (typically labeled Ha, or sometimes H1). By tradition, the null hypothesis corresponds to no change. A Type I error is **rejecting the null hypothesis when the null hypothesis is true**.

14. Definitions: Type I Error

- **Example:** Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type I error would be **allowing an ineffective drug onto the market**.

15. Definitions: Type II Error

- A Type II error is **accepting the null hypothesis when the null hypothesis is false**. **Many studies have small sample sizes that make it difficult to reject the null hypothesis,** even when there is a big change in the data. In these situations, a Type II error might be a possible explanation for the negative study results.

16. Definitions: Type II Error

- **Example:** Consider a new drug that we will put on the market if we can show that it is better than a placebo. In this context, H0 would represent the hypothesis that the average improvement (or perhaps the probability of improvement) among all patients taking the new drug is equal to the average improvement (probability of improvement) among all patients taking the placebo. A Type II error would be **keeping an effective drug off the market**.

17. What is a p-value?

- A p-value is a **measure of how much evidence we have against the null hypothesis**. The null hypothesis, traditionally represented by the symbol H0, represents the hypothesis of no change or no effect. The smaller the p-value, the more evidence we have against H0.

18. What is a p-value?

- The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true. The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

19. What is a p-value?

- The p-value is also a measure of how likely we are to get a certain sample result or a result “more extreme,” assuming H0 is true. The type of hypothesis (right tailed, left tailed or two tailed) will determine what “more extreme” means.

20. What is a p-value?

- It is easiest to understand the p-value in a data set that is already at an extreme. Suppose that a drug company alleges that **only 50% of all patients who take a certain drug will have an adverse event of some kind**. You believe that the adverse event rate is much higher. **In a sample of 12 patients, all twelve have an adverse event**.

21. What is a p-value?

- **The data supports your belief because it is inconsistent with the assumption of a 50% adverse event rate**. It would be like flipping a coin 12 times and getting heads each time.

22. What is a p-value?

- The p-value, the probability of getting a sample result of **12 adverse events in 12 patients** assuming that the adverse event rate is 50%, is a measure of this inconsistency. **The p-value, 0.000244, is small enough that we would reject the hypothesis that the adverse event rate was only 50%.**

23. What is a p-value?

A large p-value should not automatically be construed as evidence in support of the null hypothesis. Perhaps the failure to reject the null hypothesis was caused by an inadequate sample size. When you see a large p-value in a research study, you should also look for one of two things:

1. a **power calculation** that confirms that the sample size in that study was adequate for detecting a clinically relevant difference; and/or
2. a **confidence interval** that lies entirely within the range of clinical indifference.

24. What is a p-value?

You should also be cautious about a small p-value, but for different reasons. **In some situations, the sample size is so large that even differences that are trivial from a medical perspective can still achieve statistical significance**.

25. Practice exercise: interpret the p-values shown below. **The Outcome of Extubation Failure in a Community Hospital Intensive Care Unit: A Cohort Study**. Seymour CW, Martinez A, Christie JD, Fuchs BD. *Critical Care 2004*, 8:R322-R327 (20 July 2004) **Introduction:** Extubation failure has been associated with poor intensive care unit (ICU) and hospital outcomes in tertiary care medical centers. Given the large proportion of critical care delivered in the community setting, our purpose was to determine the impact of extubation failure on patient outcomes in a community hospital ICU. **Methods:** A retrospective cohort study was performed using data gathered in a 16-bed medical/surgical ICU in a community hospital. During 30 months, all patients with acute respiratory failure admitted to the ICU were included in the source population if they were mechanically ventilated by endotracheal tube for more than 12 hours. Extubation failure was defined as reinstitution of mechanical ventilation within 72 hours (n = 60), and the control cohort included patients who were successfully extubated at 72 hours (n = 93). **Results:** The primary outcome was total ICU length of stay after the initial extubation. Secondary outcomes were total hospital length of stay after the initial extubation, ICU mortality, hospital mortality, and total hospital cost. Patient groups were similar in terms of age, sex, and severity of illness, as assessed using admission Acute Physiology and Chronic Health Evaluation II score ($P > 0.05$). Both ICU (1.0 versus 10 days; $P < 0.01$) and hospital length of stay (6.0 versus 17 days; $P < 0.01$) after initial extubation were significantly longer in reintubated patients. ICU mortality was significantly higher in patients who failed extubation (odds ratio = 12.2, 95% confidence interval [CI] = 1.5–101; $P < 0.05$), but there was no significant difference in hospital mortality (odds ratio = 2.1, 95% CI = 0.8–5.4; $P < 0.15$). Total hospital costs (estimated from direct and indirect charges) were significantly increased by a mean of US$33,926 (95% CI = US$22,573–45,280; $P < 0.01$). **Conclusion:** Extubation failure in a community hospital is univariately associated with prolonged inpatient care and significantly increased cost. Corroborating data from tertiary care centers, these adverse outcomes highlight the importance of accurate predictors of extubation outcome.

26. Practice exercise: interpret the p-values shown below. **Elevated White Cell Count in Acute Coronary Syndromes: Relationship to Variants in Inflammatory and Thrombotic Genes**. Byrne CE, Fitzgerald A, Cannon CP, Fitzgerald DJ, Shields DC. *BMC Medical Genetics 2004*, 5:13 (1 June 2004) **Background:** Elevated white blood cell counts (WBC) in acute coronary syndromes (ACS) increase the risk of recurrent events, but it is not known if this is exacerbated by pro-inflammatory factors. We sought to identify whether pro-inflammatory genetic variants contributed to alterations in WBC and C-reactive protein (CRP) in an ACS population. **Methods:** WBC and genotype of interleukin 6 (IL-6 G-174C) and of interleukin-1 receptor antagonist (IL1RN intronic repeat polymorphism) were investigated in 732 Caucasian patients with ACS in the OPUS-TIMI-16 trial. Samples for measurement of WBC and inflammatory factors were taken at baseline, i.e. Within 72 hours of an acute myocardial infarction or an unstable angina event. **Results:** An increased white blood cell count (WBC) was associated with an increased C-reactive protein ($r = 0.23$, $p < 0.001$) and there was also a positive correlation between levels of β-fibrinogen and C-reactive protein ($r = 0.42$, $p < 0.0001$). IL1RN and IL6 genotypes had no significant impact upon WBC. The difference in median WBC between the two homozygote IL6 genotypes was 0.21/mm3 (95% CI = -0.41, 0.77), and -0.03/mm3 (95% CI = -0.55, 0.86) for IL1RN. Moreover, the composite endpoint was not significantly affected by an interaction between WBC and the IL1 ($p = 0.61$) or IL6 ($p = 0.48$) genotype. **Conclusions:** Cytokine pro-inflammatory genetic variants do not influence the increased inflammatory profile of ACS patients.

27. Practice exercise: interpret the p-values shown below. **Is There a Clinically Significant Gender Bias in Post-Myocardial Infarction Pharmacological Management in the Older (>60) Population of a Primary Care Practice?** Di Cecco R, Patel U, Upshur REG. *BMC Family Practice 2002*, 3:8 (3 May 2002) **Background:** Differences in the management of coronary artery disease between men and women have been reported in the literature. There are few studies of potential inequalities of treatment that arise from a primary care context. This study investigated the existence of such inequalities in the medical management of post myocardial infarction in older patients. **Methods:** A comprehensive chart audit was conducted of 142 men and 81 women in an academic primary care practice. Variables were extracted on demographic variables, cardiovascular risk factors, medical and non-medical management of myocardial infarction. **Results:** Women were older than men. The groups were comparable in terms of cardiac risk factors. A statistically significant difference (14.6%: 95% CI 0.048–28.7 p = 0.047) was found between men and women for the prescription of lipid lowering medications. 25.3% (p = 0.0005, CI 11.45, 39.65) more men than women had undergone angiography, and 14.4 % (p = 0.029, CI 2.2, 26.6) more men than women had undergone coronary artery bypass graft surgery. **Conclusion:** Women are less likely than men to receive lipid-lowering medication which may indicate less aggressive secondary prevention in the primary care setting.

28. **What is a confidence interval?**

- We statisticians have a habit of **hedging our bets**. We always insert qualifiers into our reports, warn about all sorts of assumptions, and never admit to anything more extreme than probable. There's a famous saying: **"Statistics means never having to say you're certain."**

29. **What is a confidence interval?**

- We qualify our statements, of course, because we are always **dealing with imperfect information**. In particular, we are often asked to make statements about a population (a large group of subjects) using information from a sample (a small, but carefully selected subset of this population). No matter how carefully this sample is selected to be a fair and unbiased representation of the population, **relying on information from a sample will always lead to some level of uncertainty**.

30. **What is a confidence interval?**

- **A confidence interval is a range of values that tries to quantify uncertainty associated with the sampling process**. Consider it as a **range of plausible values**.

31. **What is a confidence interval?**

- A wide confidence interval implies poor precision; we can only specify plausible values to a broad and uninformative range. A narrow confidence interval implies good precision; we can place sharp limits on the range of plausible values.

32. **What is a confidence interval?**

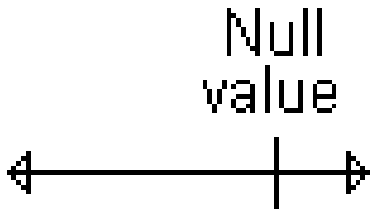
- Consider a recent study of **homoeopathic treatment of pain and swelling after oral surgery** (Lokken 1995). When examining swelling 3 days after the operation, they showed that **homoeopathy led to 1 mm less swelling on average**. The **95% confidence interval, however, ranged from -5.5 to 7.5 mm**. This interval implies that **neither a large improvement due to homoeopathy nor a large decrement could be ruled out**.

33. **What is a confidence interval?**

- When you see a confidence interval in a published medical report, you should look for two things. First, **does the interval contain a value that implies no change or no effect**? For example, with a confidence interval for a difference look to see whether that interval includes zero. With a confidence interval for a ratio, look to see whether that interval contains one.

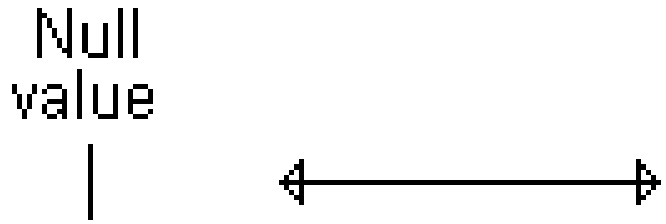
34. **What is a confidence interval?**

- Here's an example of a confidence interval that contains the null value. **The interval shown below implies no statistically significant change**.



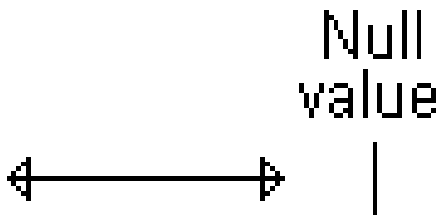
35. **What is a confidence interval?**

- Here's an example of a confidence interval that excludes the null value. If we assume that larger implies better, then **the interval shown below would imply a statistically significant improvement**.



36. **What is a confidence interval?**

- Here's a different example of a confidence interval that excludes the null value. **The interval shown below implies a statistically significant decline**.

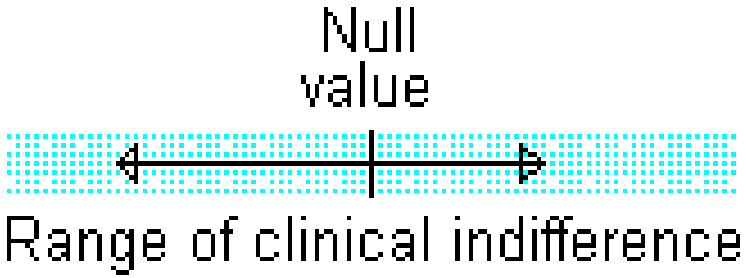


37. **What is a confidence interval?**

- You should also see **whether the confidence interval lies partly or entirely within a range of clinical indifference**. Clinical indifference represents values of such a trivial size that you would not want to change your current practice.

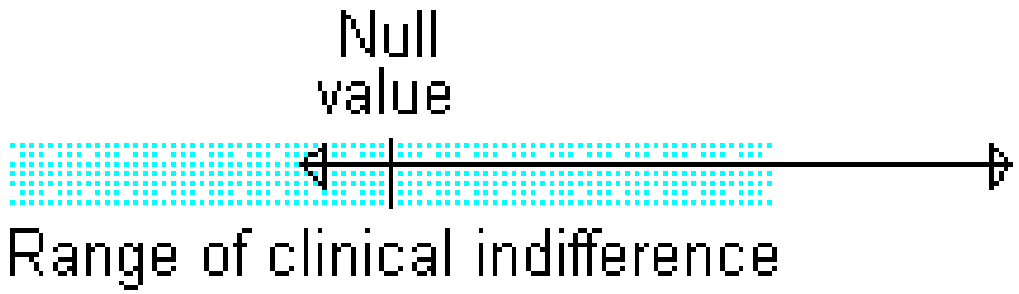
38. **What is a confidence interval?**

- **If a confidence interval is contained entirely within your range of clinical indifference**, then you have clear and convincing evidence to **keep doing things the same way**.



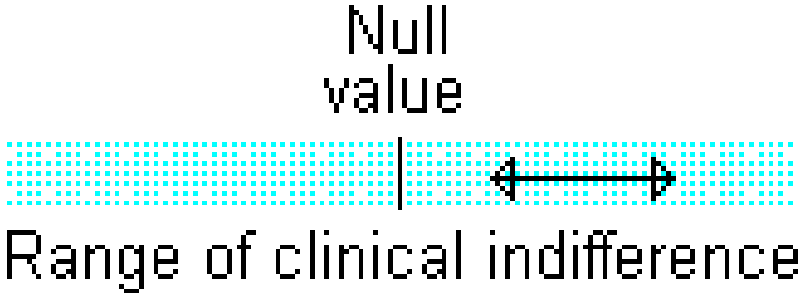
39. **What is a confidence interval?**

- One the other hand, **if part of the confidence interval lies outside the range of clinical indifference**, then you should consider the possibility that **the sample size is too small**.



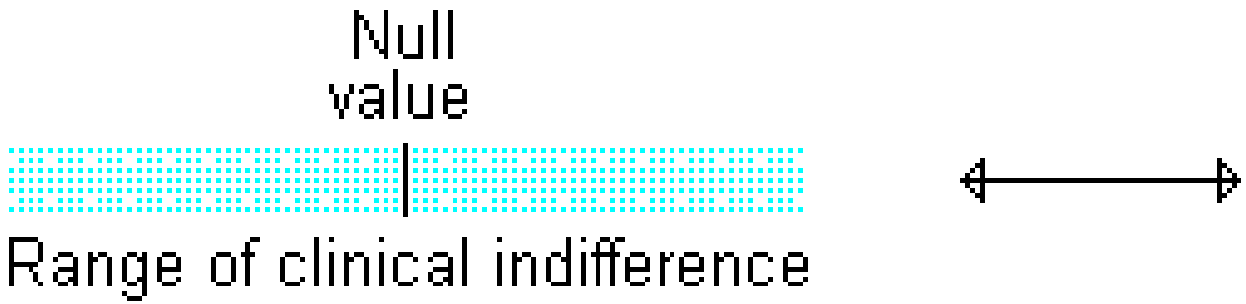
40. **What is a confidence interval?**

- If your **confidence interval excludes the null value but still lies entirely within the range of clinical indifference**, then you have a result with **statistical significance, but no practical significance**.



41. **What is a confidence interval?**

- Finally, if your **confidence interval excludes the null value and lies outside the range of clinical indifference**, then you have **both statistical and practical significance**.



42. Practice exercise: interpret the confidence intervals shown below. **The Outcome of Extubation Failure in a Community Hospital Intensive Care Unit: A Cohort Study**. Seymour CW, Martinez A, Christie JD, Fuchs BD. *Critical Care 2004*, 8:R322-R327 (20 July 2004) **Introduction:** Extubation failure has been associated with poor intensive care unit (ICU) and hospital outcomes in tertiary care medical centers. Given the large proportion of critical care delivered in the community setting, our purpose was to determine the impact of extubation failure on patient outcomes in a community hospital ICU. **Methods:** A retrospective cohort study was performed using data gathered in a 16-bed medical/surgical ICU in a community hospital. During 30 months, all patients with acute respiratory failure admitted to the ICU were included in the source population if they were mechanically ventilated by endotracheal tube for more than 12 hours. Extubation failure was defined as reinstitution of mechanical ventilation within 72 hours (n = 60), and the control cohort included patients who were successfully extubated at 72 hours (n = 93). **Results:** The primary outcome was total ICU length of stay after the initial extubation. Secondary outcomes were total hospital length of stay after the initial extubation, ICU mortality, hospital mortality, and total hospital cost. Patient groups were similar in terms of age, sex, and severity of illness, as assessed using admission Acute Physiology and Chronic Health Evaluation II score ($P > 0.05$). Both ICU (1.0 versus 10 days; $P < 0.01$) and hospital length of stay (6.0 versus 17 days; $P < 0.01$) after initial extubation were significantly longer in reintubated patients. ICU mortality was significantly higher in patients who failed extubation (odds ratio = 12.2, 95% confidence interval [CI] = 1.5–101; $P < 0.05$), but there was no significant difference in hospital mortality (odds ratio = 2.1, 95% CI = 0.8–5.4; $P < 0.15$). Total hospital costs (estimated from direct and indirect charges) were significantly increased by a mean of US$33,926 (95% CI = US$22,573–45,280; $P < 0.01$). **Conclusion:** Extubation failure in a community hospital is univariately associated with prolonged inpatient care and significantly increased cost. Corroborating data from tertiary care centers, these adverse outcomes highlight the importance of accurate predictors of extubation outcome.

43. Practice exercise: interpret the confidence intervals shown below.

Elevated White Cell Count in Acute Coronary Syndromes: Relationship to Variants in Inflammatory and Thrombotic Genes. Byrne CE, Fitzgerald A, Cannon CP, Fitzgerald DJ, Shields DC. *BMC Medical Genetics 2004*, 5:13 (1 June 2004) **Background:** Elevated white blood cell counts (WBC) in acute coronary syndromes (ACS) increase the risk of recurrent events, but it is not known if this is exacerbated by pro-inflammatory factors. We sought to identify whether pro-inflammatory genetic variants contributed to alterations in WBC and C-reactive protein (CRP) in an ACS population. **Methods:** WBC and genotype of interleukin 6 (IL-6 G-174C) and of interleukin-1 receptor antagonist (IL1RN intronic repeat polymorphism) were investigated in 732 Caucasian patients with ACS in the OPUS-TIMI-16 trial. Samples for measurement of WBC and inflammatory factors were taken at baseline, i.e. Within 72 hours of an acute myocardial infarction or an unstable angina event. **Results:** An increased white blood cell count (WBC) was associated with an increased C-reactive protein ($r = 0.23$, $p < 0.001$) and there was also a positive correlation between levels of β-fibrinogen and C-reactive protein ($r = 0.42$, $p < 0.0001$). IL1RN and IL6 genotypes had no significant impact upon WBC. The difference in median WBC between the two homozygote IL6 genotypes was 0.21/mm3 (95% CI = -0.41, 0.77), and -0.03/mm3 (95% CI = -0.55, 0.86) for IL1RN. Moreover, the composite endpoint was not significantly affected by an interaction between WBC and the IL1 ($p = 0.61$) or IL6 ($p = 0.48$) genotype. **Conclusions:** Cytokine pro-inflammatory genetic variants do not influence the increased inflammatory profile of ACS patients.

44. Practice exercise: interpret the confidence intervals shown below. **Is There a Clinically Significant Gender Bias in Post-Myocardial Infarction Pharmacological Management in the Older (>60) Population of a Primary Care Practice?** Di Cecco R, Patel U, Upshur REG. *BMC Family Practice 2002*, 3:8 (3 May 2002) **Background:** Differences in the management of coronary artery disease between men and women have been reported in the literature. There are few studies of potential inequalities of treatment that arise from a primary care context. This study investigated the existence of such inequalities in the medical management of post myocardial infarction in older patients. **Methods:** A comprehensive chart audit was conducted of 142 men and 81 women in an academic primary care practice. Variables were extracted on demographic variables, cardiovascular risk factors, medical and non-medical management of myocardial infarction. **Results:** Women were older than men. The groups were comparable in terms of cardiac risk factors. A statistically significant difference (14.6%: 95% CI 0.048–28.7 p = 0.047) was found between men and women for the prescription of lipid lowering medications. 25.3% (p = 0.0005, CI 11.45, 39.65) more men than women had undergone angiography, and 14.4 % (p = 0.029, CI 2.2, 26.6) more men than women had undergone coronary artery bypass graft surgery. **Conclusion:** Women are less likely than men to receive lipid-lowering medication which may indicate less aggressive secondary prevention in the primary care setting.

45. Repeat of pop quiz #1

A research paper computes a p-value of 0.45. How would you interpret this p-value?

1. Strong evidence for the null hypothesis
2. Strong evidence for the alternative hypothesis
3. Little or no evidence for the null hypothesis
4. Little or no evidence for the alternative hypothesis
5. More than one answer above is correct.
6. I do not know the answer.

46. Repeat of pop quiz #2

A research paper computes a confidence interval for a relative risk of 0.82 to 3.94. What does this confidence interval tell you.

1. The result is statistically significant and clinically important.
2. The result is not statistically significant, but is clinically important.
3. The result is statistically significant, but not clinically important.
4. The result is not statistically significant, and not clinically important.
5. The result is ambiguous.
6. I do not know the answer.

47. How bad is it, really? Measures of risk.

- Abstract: The odds ratio and the relative risk are both measures of risk used for binary outcomes, but sometimes they can differ markedly from one another. The relative risk offers a more natural interpretation, but certain research designs preclude its computation.

48. Objectives

In this class you will learn how to:

- compute an odds ratio and a relative risk from a two by two table;
- list the types of research designs where the relative risk should not be computed, and

49. Sources

Part of the material for this webinar comes from:

- Simon SD. Understanding the odds ratio and the relative risk. J Androl. 2001 Jul-Aug;22(4):533-6.
- Stats: Odds ratio versus relative risk (January 9, 2001).
  - http://www.childrens-mercy.org/stats/journal/oddsratio.asp

50. Very bad joke

A doctor is advising her patient about the risks of an upcoming surgery. She warned that the probability that the patient would die during surgery was 60%. Then she looked up an said, no wait, the risk is twice as big in your demographic group. The chances that you will die during surgery is actually 120%. The patient seemed a bit confused. I know what a 100% risk of mortality would be—I'm a goner. But what would a 120% risk of mortality be? The doctor replied, that is a fate worse than death.

51. Pop quiz #3

A relative risk should not be computed for the following design because the prevalence of the disease is artificially constrained.

1. Case-control design
2. Cohort design
3. Cross-sectional design
4. Historical control design
5. Don’t know/Not sure

52. Pop quiz #4

The odds ratio and the relative risk are close to one another when

1. The prevalence of the disease is low
2. The prevalence of the disease is high
3. The sample size is small
4. The sample size is large
5. Don’t know/Not sure

53. What are odds?

If you head south from Kansas City on Highway 71, you will encounter a town called “Peculiar”. This town is very proud of its name and has a sign which says “Welcome to Peculiar, where the odds are with you.”

Mathematicians and gamblers use odds frequently but the concept may be alien to most of the rest of the public. Odds is the ratio of successes to failures.

54. What are odds?

“If there is a 50-50 chance that something will go wrong, then nine times out of ten it will.” (Paul Harvey).

In this silly example a 50-50 chance means one success for every failure or 1 to 1 odds. This is sometimes called even odds.

Nine times out of ten means one success for every nine failures or one to nine odds.

55. What are odds?

To be perfectly accurate, you should specify whether you are talking about the odds of success or the odds of failure, but in most setting, it should be obvious from the context.

If your odds of winning the lottery are a million to one, that means either that:

- One million people win for every person that loses, or
- One person winds for every million that lose.

56. What are odds?

If you know the probability of a success, you can calculate the odds using the formula

- Odds = prob / (1- prob).
- For example, a probability of 0.25 corresponds to an odds of 0.25 / (1-0.25) = 0.25 / 0.75 = 1 / 3. This means that for every single success, there are three failures.

If you know the odds, then you can calculate the probability of success using the formula

- Prob = Odds / (1 + Odds).
- For example, if the odds are 3 to 1, then prob = 3 / (1 + 3) = 3 / 4.

57. Odds ratio/relative risk

Consider the following data on survival of passengers on the Titanic. Clearly, a male passenger on the Titanic was more likely to die than a female passenger. But how much more likely? You can compute the odds ratio or the relative risk to answer this question.

	Alive	Dead	Total
Female	308	154	462
Male	142	709	851
Total	450	863	1313

58. Odds ratio/relative risk

The odds ratio compares the relative odds of death in each group.

- For females, 2 to 1 odds against dying
- For males, almost 5 to 1 in favor of death

The odds ratio is approximately 10.

	Alive	Dead	Odds	
Female	308	154	154/308 = 0.5	(2 to 1 against)
Male	142	709	706/142 = 4.993	(5 to 1 in favor)

Odds ratio=4.993/0.5
=9.986

59. Odds ratio/relative risk

The relative risk (sometimes called the risk ratio) compares the probability of death in each group rather than the odds.

- The females probability of death is 1/3 (2/6).
- The male probability of death is 5/6.

The relative risk of death is 2.5

	Dead	Total	Probability	
Female	154	462	154/462 = 0.3333	(1/3 chance)
Male	709	851	709/851 = 0.8331	(5/6 chance)

Relative risk

0.8331/0.3333 = 2.5

60. Odds ratio/relative risk

There is quite a difference. Both measurements show that men were more likely to die. But the odds ratio implies that men are much worse off than the relative risk. Which number is a fairer comparison?

61. Odds ratio/relative risk

There are three issues here:

1. The relative risk measures events in a way that is interpretable and consistent with the way people really think.
2. The relative risk, though, cannot always be computed in a research design.
3. Also, the relative risk can sometimes lead to ambiguous and confusing situations.

But first, let's practice calculating some odds ratios and relative risks.

62. Practice exercise.

Read the abstract from **Socioeconomic disparities in intimate partner violence against Native American women: a cross-sectional study.** Malcoe LH, Duran BM, Montgomery JM. BMC Med 2004: 2(1); 20. The authors report an adjusted odds ratio of 5.0 for low socioeconomic index. Compute a crude odds ratio using the data that appears in the abstract. Does it differ much from the adjusted odds ratio? Interpret the adjusted odds ratio and its associated confidence interval.

63. Practice exercise. **BACKGROUND:** Intimate partner violence (IPV) against women is a global public health problem, yet data on IPV against Native American women are extremely limited. We conducted a cross-sectional study of Native American women to determine prevalence of lifetime and past-year IPV and partner injury; examine IPV in relation to pregnancy; and assess demographic and socioeconomic correlates of past-year IPV.
METHODS: Participants were recruited from a tribally-operated clinic serving low-income pregnant and childbearing women in southwest Oklahoma. A self-administered survey was completed by 312 Native American women (96% response rate) attending the clinic from June through August 1997. Lifetime and past-year IPV were measured using modified 18-item Conflict Tactics Scales. A socioeconomic index was created based on partner's education, public assistance receipt, and poverty level.
RESULTS: More than half (58.7%) of participants reported lifetime physical and/or sexual IPV; 39.1% experienced severe physical IPV; 12.2% reported partner-forced sexual activity; and 40.1% reported lifetime partner-perpetrated injuries. A total of 273 women had a spouse or boyfriend during the previous 12 months (although all participants were Native American, 59.0% of partners were non-Native). Among these women, past-year prevalence was 30.1% for physical and/or sexual IPV; 15.8% for severe physical IPV; 3.3% for forced partner-perpetrated sexual activity; and 16.4% for intimate partner injury. Reported IPV prevalence during pregnancy was 9.3%. Pregnancy was not associated with past-year IPV (odds ratio = 0.9). Past-year IPV prevalence was 42.8% among women scoring low on the socioeconomic index, compared with 10.1% among the reference group. After adjusting for age, relationship status, and household size, low socioeconomic index remained strongly associated with past-year IPV (odds ratio = 5.0; 95% confidence interval: 2.4, 10.7).
CONCLUSIONS: Native American women in our sample experienced exceptionally high rates of lifetime and past-year IPV. Additionally, within this low-income sample, there was strong evidence of socioeconomic variability in IPV. Further research should determine prevalence of IPV against Native American women from diverse tribes and regions, and examine pathways through which socioeconomic disadvantage may increase their IPV risk.

64. Practice exercise (hints)

1. A total of 273 women had a spouse or boyfriend during the previous 12 months.
2. Among these women, past-year prevalence was 30.1% for physical and/or sexual IPV.
3. Past-year IPV prevalence was 42.8% among women scoring low on the socioeconomic index
4. compared with 10.1% among the reference group

	IPV	No	Total
Low SES	42.8% of a[3]		a
Reference group	10.1% of b[4]		b
Total	30.3% of 273[2]		273[1]

65. Practice exercise (hints)

You don’t need to know the row totals (a and b) in order to calculate odds.

	IPV	No	Odds
Low SES	42.8% of a	57.4% of a	57.4/42.8 = 1.34
Reference group	10.1% of b	89.9% of b	89.9/10.1 = 8.9

Odds ratio

8.9/1.34 = 6.64

66. Practice exercise

Read the abstract from **Tongue lesions in psoriasis: a controlled study.** Daneshpazhooh M, Moslehi H, Akhyani M, Etesami M. BMC Dermatol 2004: 4(1); 16. The crude odds ratios for Fissured Tongue and for benign migratory glossitis have been removed from this abstract. Calculate these value using the information provided in the abstract. Interpret these odds ratios and the associated confidence intervals.

67. **BACKGROUND:** Our objective was to study tongue lesions and their significance in psoriatic patients.

METHODS: The oral mucosa was examined in 200 psoriatic patients presenting to Razi Hospital in Tehran, Iran, and 200 matched controls.

RESULTS: Fissured tongue (FT) and benign migratory glossitis (BMG) were the two most frequent findings. FT was seen more frequently in psoriatic patients (n = 66, 33%) than the control group (n = 19, 9.5%) [odds ratio (OR): **[DELETED]**; 95% confidence interval (CI): 2.61-8.52] (p-value < 0.0001). BMG, too, was significantly more frequent in psoriatic patients (28 cases, 14%) than the control group (12 cases, 6%) (OR: **[DELETED]**; 95% CI: 1.20-5.50) (p-value < 0.012). In 11 patients (5.5%), FT and BMG coexisted. FT was more frequent in pustular psoriasis (7 cases, 53.8%) than erythemato-squamous types (56 cases, 30.4%). On the other hand, the frequency of BMG increased with the severity of psoriasis in plaque-type psoriasis assessed by psoriasis area and severity index (PASI) score.

CONCLUSIONS: Nonspecific tongue lesions are frequently observed in psoriasis. Further studies are recommended to substantiate the clinical significance of these seemingly nonspecific findings in suspected psoriatic cases.

68. Practice exercise (hints)

To help you get started, note that

1. The oral mucosa was examined in 200 psoriatic patients
2. and 200 matched controls
3. FT was seen more frequently in psoriatic patients (n = 66, 33%)
4. than the control group (n = 19, 9.5%)

	Fissured tongue	No	Total
Psoriatic	66(3)		200(1)
Control	19(4)		200(2)
Total			

69. Practice exercise

Read the abstract from **Breastfeeding practices in a cohort of inner-city women: the role of contraindications.** England L, Brenner R, Bhaskar B, Simons-Morton B, Das A, Revenis M, Mehta N, Clemens J. BMC Public Health 2003: 3(1); 28.. The authors report an adjusted odds ratio of 0.19 for presence of contraindication. Compute a crude odds ratio using the data that appears in the abstract. Does it differ much from the adjusted odds ratio? Interpret the adjusted odds ratio and its associated confidence interval.

70. **BACKGROUND:** Little is known about the role of breastfeeding contraindications in breastfeeding practices. Our objectives were to 1) identify predictors of breastfeeding initiation and duration among a cohort of predominantly low-income, inner-city women, and 2) evaluate the contribution of breastfeeding contraindications to breastfeeding practices.
METHODS: Mother-infant dyads were systematically selected from 3 District of Columbia hospitals between 1995 and 1996. Breastfeeding contraindications and potential predictors of breastfeeding practices were identified through medical record reviews and interviews conducted after delivery (baseline). Interviews were conducted at 3-7 months postpartum and again at 7-12 months postpartum to determine breastfeeding initiation rates and duration. Multivariable logistic regression analysis was used to identify baseline factors associated with initiation of breastfeeding. Cox proportional hazards models were generated to identify baseline factors associated with duration of breastfeeding.
RESULTS: Of 393 study participants, 201 (51%) initiated breastfeeding. A total of 61 women (16%) had at lease one documented contraindication to breastfeeding; 94% of these had a history of HIV infection and/or cocaine use. Of the 332 women with no documented contraindications, 58% initiated breastfeeding, vs. 13% of women with a contraindication. In adjusted analysis, factors most strongly associated with breastfeeding initiation were presence of a contraindication (adjusted odds ratio [AOR], 0.19; 95% confidence interval [CI], 0.08-0.47), and mother foreign-born (AOR, 4.90; 95% CI, 2.38-10.10). Twenty-five percent of study participants who did not initiate breastfeeding cited concern about passing dangerous things to their infants through breast milk. Factors associated with discontinuation of breastfeeding (all protective) included mother foreign-born (hazard ratio [HR], 0.55; 95% CI 0.39-0.77) increasing maternal age (HR for 5-year increments, 0.80; 95% CI, 0.69-0.92), and infant birth weight > or = 2500 grams (HR, 0.45; 95% CI, 0.26-0.80).
CONCLUSIONS: Breastfeeding initiation rates and duration were suboptimal in this inner-city population. Many women who did not breastfeed had contraindications and/or were concerned about passing dangerous things to their infants through breast milk. It is important to consider the prevalence of contraindications to breastfeeding when evaluating breastfeeding practices in high-risk communities.

71. Practice exercise (hints)

Here's something to help you get started.

1. Of 393 study participants,
2. 201 (51%) initiated breastfeeding.
3. A total of 61 women (16%) had at lease one documented contraindication to breastfeeding;
4. Of the 332 women with no documented contraindications,
5. 58% initiated breastfeeding,
6. vs. 13% of women with a contraindication.

	Initiate BF	No	Total
Contraindication	13% of 61 (6)		61 (3)
No	58% of 332 (5)		332 (4)
Total	201 (2)		393 (1)

72. Practice exercise

Read the abstract from **Treatment of Retinopathy of Prematurity with topical ketorolac tromethamine: a preliminary study.** Avila-Vazquez M, Maffrand R, Sosa M, Franco M, De Alvarez BV, Cafferata ML, Bergel E. BMC Pediatr 2004: 4(1); 15. The relative risk for cryotherapy has been removed. Calculate this value using the information provided in the abstract. Interpret this relative risk and the associated confidence interval.

73. **BACKGROUND:** Retinopathy of Prematurity (ROP) is a common retinal neovascular disorder of premature infants. It is of variable severity, usually heals with mild or no sequelae, but may progress to blindness from retinal detachments or severe retinal scar formation. This is a preliminary report of the effectiveness and safety of a new and original use of topical ketorolac in preterm newborn to prevent the progression of ROP to the more severe forms of this disease.
METHODS: From January 2001 to December 2002, all fifty nine preterm newborns with birthweight less than 1250 grams or gestational age less than 30 weeks of gestational age admitted to neonatal intensive care were eligible for treatment with topical ketorolac (0.25 milligrams every 8 hours in each eye). The historical comparison group included all 53 preterm newborns, with the same inclusion criteria, admitted between January 1999 and December 2000.
RESULTS: Groups were comparable in terms of weight distribution, Apgar score at 5 minutes, incidence of sepsis, intraventricular hemorrhage and necrotizing enterocolitis. The duration of oxygen therapy was significantly longer in the control group. In the ketorolac group, among 43 children that were alive at discharge, one (2.3%) developed threshold ROP and cryotherapy was necessary. In the comparison group 35 children survived, and six child (17%) needed cryotherapy (Relative Risk **[DELETED]**, 95%CI 0.00 to 0.80, p = 0.041). Adjusting by duration of oxygen therapy did not significantly change these results. Adverse effects attributable to ketorolac were not detected.
CONCLUSIONS: This preliminary report suggests that ketorolac in the form of an ophthalmic solution can reduce the risk of developing severe ROP in very preterm newborns, without producing significant adverse side effects. These results, although promising, should be interpreted with caution because of the weakness of the study design. This is an inexpensive and simple intervention that might ameliorate the progression of a disease with devastating consequences for children and their families. We believe that next logical step would be to assess the effectiveness of this intervention in a randomized controlled trial of adequate sample size.

74. Practice exercise (hints)

Here's something to help you get started.

1. In the ketorolac group, among 43 children that were alive at discharge,
2. one (2.3%) developed threshold ROP and cryotherapy was necessary.
3. In the comparison group 35 children survived,
4. and six child (17%) needed cryotherapy

	ROP	No	Total
Ketorolac	1 (2.3%)[(2)]		43[(1)]
Control	6 (17%)[(4)]		35[(3)]
Total			

75. Practice exercise

Read the abstract from **Misoprostol for treating postpartum haemorrhage: a randomized controlled trial [ISRCTN72263357].** Hofmeyr GJ, Ferreira S, Nikodem VC, Mangesi L, Singata M, Jafta Z, Maholwana B, Mlokoti Z, Walraven G, Gulmezoglu AM. BMC Pregnancy Childbirth 2004: 4(1); 16. The relative risks for reduced blood loss, shivering, and pyrexia have been removed. Calculate these values using the information provided in the abstract. Interpret these relative risks and their associated confidence intervals.

76. **BACKGROUND:** Postpartum haemorrhage remains an important cause of maternal death despite treatment with conventional therapy. Uncontrolled studies and one randomised comparison with conventional oxytocics have reported dramatic effects with high-dose misoprostol, usually given rectally, for treatment of postpartum haemorrhage, but this has not been evaluated in a placebo-controlled trial.
METHODS: The study was conducted at East London Hospital Complex, Tembisa and Chris Hani Baragwanath Hospitals, South Africa. Routine active management of the third stage of labour was practised. Women with more than usual postpartum bleeding thought to be related to inadequate uterine contraction were invited to participate, and to sign informed consent. All routine treatment was given from a special 'Postpartum Haemorrhage Trolley'. In addition, participants who consented were enrolled by drawing the next in a series of randomised treatment packs containing either misoprostol 5 x 200 microg or similar placebo, which were given 1 orally, 2 sublingually and 2 rectally.
RESULTS: With misoprostol there was a trend to reduced blood loss >/=500 ml in 1 hour after enrolment measured in a flat plastic 'fracture bedpan', the primary outcome (6/117 vs 11/120, relative risk **[DELETED]**; 95% confidence interval 0.21 to 1.46). There was no difference in mean blood loss or haemoglobin level on day 1 after birth < 6 g/dl or blood transfusion. Side-effects were increased, namely shivering (63/116 vs 30/118; **[DELETED]**, 1.50 to 3.04) and pyrexia > 38.5 degrees C (11/114 vs 2/118; **[DELETED]**, 1.29 to 25). In the misoprostol group 3 women underwent hysterectomy of whom 1 died, and there were 2 further maternal deaths.
CONCLUSIONS: Because of a lower than expected incidence of the primary outcome in the placebo group, the study was underpowered. We could not confirm the dramatic effect of misoprostol reported in several unblinded studies, but the results do not exclude a clinically important effect. Larger studies are needed to assess substantive outcomes and risks before misoprostol enters routine use.

77. Review of major points

1. The relative risk has a more natural interpretation than the odds ratio.
2. You should not use the relative risk for certain research designs where the prevalence is artificially constrained.

78. Repeat of pop quiz #3

A relative risk should not be computed for the following design because the prevalence of the disease is artificially constrained.

1. Case-control design
2. Cohort design
3. Cross-sectional design
4. Historical control design
5. Don’t know/Not sure

79. Repeat of pop quiz #4

The odds ratio and the relative risk are close to one another when

1. The prevalence of the disease is low
2. The prevalence of the disease is high
3. The sample size is small
4. The sample size is large
5. Don’t know/Not sure

80. Putting your life in the hands of a coin: Randomized trials.

- Abstract: In research studies that compare a treatment group and a control group, you need to assess whether the comparison is a fair comparison—an apples to apples comparison. Randomization is a simple method that insures that patients assigned to the treatment group are comparable to patients assigned to the control group. There are, however, practical and ethical constraints that can sometimes prevent the use of randomization.

81. Objectives

In this class you will learn how to:

- describe how covariate imbalance can threaten the validity of a research study,
- explain how randomization prevents covariate imbalance, and
- understand the practical and ethical limitations to randomized studies.

82. Sources

Part of the material for this webinar comes from:

- Simon SD. Statistical Evidence in Medical Trials, What Do the Data Really Tell Us? 2006. Oxford University Press: Oxford, England.
- Simon SD. Is the randomized clinical trial the gold standard of research?. J Androl. 2001 Nov-Dec;22(6):938-43.
- Stats #32a: Statistical Evidence: Apples or Oranges? Randomized studies.
  - http://www.childrens-mercy.org/stats/training/hand32a.asp

83. Pop quiz #5

When the demographic profile of the patients in your treatment group differ sharply from the profile of patients in your control group, you have:

1. covariate imbalance,
2. observational data,
3. response bias,
4. spectrum bias,
5. stratified data,
6. don’t know/not sure

84. Pop quiz #6

Randomization is not practical:

1. when doctors believe that the new treatment is superior to the current standard
2. when patients have a strong preference for a particular treatment
3. when the experiment requires deliberate exposure of patients to something that is known to be harmful
4. randomization is impractical for all of the above situations
5. randomization can be applied easily in all of the above situations
6. don’t know/not sure

85. Pop quiz #7

The following approaches are credible alternatives to randomization:

1. alternating between treatment and control
2. assigning all new patients to the treatment group and choosing controls from a medical database
3. assigning treatment group on the basis of the last digit of your birthday
4. letting the doctor choose whether a patient gets into the treatment group or the control group
5. none of these approaches is as effective as randomization
6. don’t know/not sure

86. Covariate imbalance

Almost all research involves comparison. Do women who take Tamoxifen have a lower rate of breast cancer recurrence than women who take a placebo? Do left-handed people die at an earlier age than right-handed people? Are men with severe vertex balding more likely to develop heart disease than men with no balding?

87. Covariate imbalance

When you make a comparison between a treatment group and a control group, you want a fair comparison. You want the control group to be identical to the treatment group in all respects, except for the treatment in question. You want an apples-to-apples comparison.

88. Covariate imbalance

Sometimes, however, you get an unfair comparison, an apples-to-oranges comparison. The control group differs on some important characteristics that might influence the outcome measure. This is known as covariate imbalance. Covariate imbalance is not an insurmountable problem, but it does make a study less authoritative.

89. Covariate imbalance

Women who take oral contraceptives appear to have a higher risk of cervical cancer. But covariate imbalance might be producing an artificial rise in cancer rates for this group. Women who take oral contraceptives behave, as a group, differently than other women.

90. Covariate imbalance

For example, women who take oral contraceptives have a larger number of pap smears. This is probably because these women visit their doctors more regularly in order to get their prescriptions refilled and therefore have more opportunities to be offered a pap smear. This difference could lead to an increase in the number of detected cancer cases. Perhaps the other women have just as much cancer, but it is more likely to remain undetected.

91. Covariate imbalance

- There are many other variables that influence the development of cervical cancer: age of first intercourse, number of sexual partners, use of condoms, and smoking habits. If women who take oral contraceptives differ in any of these lifestyle factors, then that might also produce a difference in cervical cancer rates.
- The possibility that oral contraceptives causes an increase in the risk of cervical cancer is quite complex; a good summary of all the issues involved is available at:
  - www.jhuccp.org/pr/a9/a9chap5.shtml.

92. Randomization

One way to avoid most of the problems with imbalanced covariates is to use randomization. Randomization is the assignment of treatment groups through the use of a random device, like the flip of a coin or the roll of a die, or numbers randomly generated by a computer. Randomization is not always possible, practical, or ethical. But when you can use randomization, it greatly adds to the credibility of the research study.

93. Randomization

In a randomized study, the researchers have a high degree of control over the patients. They decide who gets what. This is a hallmark of a randomized design and it only can occur when the patients and/or their doctors have no say in the assignment. This is an incredible gift that patients in a research study offer you. They sacrifice their ability to choose between two therapies and instead let that choice be decided by the flip of a coin.

94. Randomization

Randomization helps ensure that both measurable and immeasurable factors are balanced out across both the standard and the new therapy, assuring a fair comparison. Used correctly, it also guarantees that no conscious or subconscious efforts were used to allocate subjects in a biased way.

95. Randomization

Randomization relies on the law of large numbers. With small sample sizes, covariate imbalance may still sneak in. A study examining the probability of covariate imbalance (Hsu 1989) showed that total sample sizes less than 10 could have a 50% chance or higher of having a categorical covariate with levels twice as large in one group than the other. This study also showed that total sample sizes of 40 or greater would have very little chance of such a serious imbalance.

96. A fishy story about randomization

I was told this story but have no way of verifying its accuracy. It is one of those stories that if it is not true, it should be. A long, long, time ago, a research group wanted to examine a pollutant to find concentration levels that would kill fish. This research required that 100 fish be separated into five tanks, each of which would get a different level of the pollutant. The researchers caught the first 20 fish and put them in the first tank, then put the next 20 fish in a second tank, and so forth. The last 20 fish went into the fifth tank. Each fish tank got a different concentration of the pollutant.

97. A fishy story about randomization

When the research was done, the mortality was related not to the dosage, but to the order in which the tanks were filled, with the worst outcomes being in the first tank filled and the best outcomes in the last tank filled. What happened was that the slow-moving, easy-to-catch fish (the weakest and most sickly) were all allocated to the first tank. The fast-moving, hard-to-catch fish (the strongest and healthiest) ended up in the last tank.

.

98. Concealed allocation

Another important aspect of randomization is concealed allocation, which is withholding the randomization list from those involved with recruiting subjects. This concealment occurs until after subjects agree to participate and the recruiter determines that the patient is eligible for the study. Only then is a sealed envelope opened that reveals the treatment status. Concealed allocation can also be done through a special phone number that the doctor calls to discover the treatment status.

99. Concealed allocation

If the randomization list is not concealed, doctors have the ability to consciously or unconsciously influence the composition of the groups. They can do this by applying exclusion criteria differentially or by delaying entry of a certain healthier (or unhealthier) subject so he/she gets into the ‘desirable’ group. Unblinded allocation schemes tend, on average to overstate the effectiveness of the new therapy by 30–40% (Schulz 1996).

100. Ethical and practical constraints on randomization

There are many situations where randomization is not practical or possible. Sometimes patients have a strong preference for one particular treatment and would not consider the possibility of being randomized into a different treatment. Surgery is one area with strong patient preferences especially for newer approaches like laparoscopic surgery (Lefering 2003).

101. Ethical and practical constraints on randomization

Randomization is also problematic for interventions that are already known to be effective. While further research would help better define these advantages, you cannot ask half of your patients to sacrifice the benefits of the new intervention.

102. Ethical and practical constraints on randomization

Randomization also does not work when you are studying noxious agents, like second-hand cigarette smoke or noisy workplaces. It would be unethical to deliberately expose people to any of these agents, so we have use non-randomized studies of people who are unfortunate enough to be trapped in settings with noxious agents.

103. Ethical and practical constraints on randomization

Sometimes researchers just do not want to go to the effort of randomizing. If you assign the treatment or therapy, rather than letting the patients and their doctors choose, you have to expend a lot of energy. Is it worth the effort? It is usually faster and cheaper to use existing nonrandomized databases. You get a lot larger sample size for your money. Depending on the situation, that might be enough to counterbalance the advantages of randomization.

104. Variations on randomization

There are three variations to randomization where the researchers have control over treatment assignment, but they use something other than a table of random numbers for the assignment. The first approach, minimization, is a credible and reasonable choice, but the other two approaches, alternating assignment and haphazard assignment, do not have much to recommend them.

105. Variations on randomization

An alternative, when the researchers have sufficient control, is to allocate the assignments so that at each step, the covariate imbalance is minimized.

So if the treatment group has a slight surplus of older patients and the next patient to join the study is also older than average, then that patient would be assigned to the control group so as to reduce the age discrepancy.

106. Variations on randomization

Another approach used in place of randomization is to alternate the assignment, so that every even patient is in the treatment group and every odd patient is in the control group. Alternating assignment was popular in trials before World War II; it was felt that researchers would not understand and not tolerate randomization (Yoshioka 1998).

107. Variations on randomization

Alternating assignment seems on the surface to be a good approach, but it can sometimes lead to trouble. This is especially true when one patient has a direct or indirect influence on the next patient. You may have seen this level of influence if you grow vegetables in a garden. If you have a row of cabbages, for example, you will often see a pattern of big cabbage, little cabbage, big cabbage, little cabbage, etc.

108. Variations on randomization

What happens, if the cabbages are planted a bit too closely, is that one of the cabbages will grow just a bit faster at first. It will extend into the neighboring cabbage’s territory, stealing some of the nutrients and water, and thus growing even faster at the expense of the neighbor. If you assigned a fertilizer to every other cabbage, you would probably see an artificial difference because of the alternating pattern in growth within a row.

109. Variations on randomization

Haphazard assignment uses some arbitrary value like a birthdate or social security number to assign patients to groups. Often it is the evenness/oddness of the arbitrary number that determines the treatment assignment. For example, patients born on even-numbered dates would be assigned to the treatment group and those born on odd-numbered dates would be assigned to the control group. An arbitrary or haphazard number is never going to be as good as a purely random number. The haphazard assignment will always cast a shadow of doubt over the research study.

110. Practice exercises

For each of the following abstracts, randomization was NOT used. Explain why it would be impractical or unethical to conduct a randomized experiment in each of these settings.

111. **Body fatness during childhood and adolescence and incidence of breast cancer in premenopausal women: a prospective cohort study**. Heather J Baer, Graham A Colditz, Bernard Rosner, Karin B Michels, Janet W Rich-Edwards, David J Hunter and Walter C Willett. Breast Cancer Research 2005, 7:R314-R325 doi:10.1186/bcr998. **Introduction** Body mass index (BMI) during adulthood is inversely related to the incidence of premenopausal breast cancer, but the role of body fatness earlier in life is less clear. We examined prospectively the relation between body fatness during childhood and adolescence and the incidence of breast cancer in premenopausal women. **Methods** Participants were 109,267 premenopausal women in the Nurses' Health Study II who recalled their body fatness at ages 5, 10 and 20 years using a validated 9-level figure drawing. Over 12 years of follow up, 1318 incident cases of breast cancer were identified. Cox proportional hazards regression was used to compute relative risks (RRs) and 95% confidence intervals (CIs) for body fatness at each age and for average childhood (ages 5–10 years) and adolescent (ages 10–20 years) fatness. **Results** Body fatness at each age was inversely associated with premenopausal breast cancer incidence; the multivariate RRs were 0.48 (95% CI 0.35–0.55) and 0.57 (95% CI 0.39–0.83) for the most overweight compared with the most lean in childhood and adolescence, respectively (P for trend < 0.0001). The association for childhood body fatness was only slightly attenuated after adjustment for later BMI, with a multivariate RR of 0.52 (95% CI 0.38–0.71) for the most overweight compared with the most lean (P for trend = 0.001). Adjustment for menstrual cycle characteristics had little impact on the association. **Conclusion** Greater body fatness during childhood and adolescence is associated with reduced incidence of premenopausal breast cancer, independent of adult BMI and menstrual cycle characteristics. http://breast-cancer-research.com/content/7/3/R314

112. **Impact of a nurses' protocol-directed weaning procedure on outcomes in patients undergoing mechanical ventilation for longer than 48 hours: a prospective cohort study with a matched historical control group**. Jean-Marie Tonnelier, Gwenaël Prat, Grégoire Le Gal, Christophe Gut-Gobert, Anne Renault, Jean-Michel Boles and Erwan L'Her. Critical Care 2005, 9:R83-R89 doi:10.1186/cc3030. **Introduction** The aim of the study was to determine whether the use of a nurses' protocol-directed weaning procedure, based on the French intensive care society (SRLF) consensus recommendations, was associated with reductions in the duration of mechanical ventilation and intensive care unit (ICU) length of stay in patients requiring more than 48 hours of mechanical ventilation. **Methods** This prospective study was conducted in a university hospital ICU from January 2002 through to February 2003. A total of 104 patients who had been ventilated for more than 48 hours and were weaned from mechanical ventilation using a nurses' protocol-directed procedure (cases) were compared with a 1:1 matched historical control group who underwent conventional physician-directed weaning (between 1999 and 2001). Duration of ventilation and length of ICU stay, rate of unsuccessful extubation and rate of ventilator-associated pneumonia were compared between cases and controls. **Results** The duration of mechanical ventilation (16.6 ± 13 days versus 22.5 ± 21 days; P = 0.02) and ICU length of stay (21.6 ± 14.3 days versus 27.6 ± 21.7 days; P = 0.02) were lower among patients who underwent the nurses' protocol-directed weaning than among control individuals. Ventilator-associated pneumonia, ventilator discontinuation failure rates and ICU mortality were similar between the two groups. **Discussion** Application of the nurses' protocol-directed weaning procedure described here is safe and promotes significant outcome benefits in patients who require more than 48 hours of mechanical ventilation.
http://ccforum.com/content/9/2/R83

113. **Extravascular lung water in patients with severe sepsis: a prospective cohort study**. Greg S Martin, Stephanie Eaton, Meredith Mealer and Marc Moss. Critical Care 2005, 9:R74-R82 doi:10.1186/cc3025. **Introduction** Few investigations have prospectively examined extravascular lung water (EVLW) in patients with severe sepsis. We sought to determine whether EVLW may contribute to lung injury in these patients by quantifying the relationship of EVLW to parameters of lung injury, to determine the effects of chronic alcohol abuse on EVLW, and to determine whether EVLW may be a useful tool in the diagnosis of acute respiratory distress syndrome (ARDS). **Methods** The present prospective cohort study was conducted in consecutive patients with severe sepsis from a medical intensive care unit in an urban university teaching hospital. In each patient, transpulmonary thermodilution was used to measure cardiovascular hemodynamics and EVLW for 7 days via an arterial catheter placed within 72 hours of meeting criteria for severe sepsis. **Results** A total of 29 patients were studied. Twenty-five of the 29 patients (86%) were mechanically ventilated, 15 of the 29 patients (52%) developed ARDS, and overall 28-day mortality was 41%. Eight out of 14 patients (57%) with non-ARDS severe sepsis had high EVLW with significantly greater hypoxemia than did those patient with low EVLW (mean arterial oxygen tension/fractional inspired oxygen ratio 230.7 ± 36.1 mmHg versus 341.2 ± 92.8 mmHg; $P < 0.001$). Four out of 15 patients with severe sepsis with ARDS maintained a low EVLW and had better 28-day survival than did ARDS patients with high EVLW (100% versus 36%; $P = 0.03$). ARDS patients with a history of chronic alcohol abuse had greater EVLW than did nonalcoholic patients (19.9 ml/kg versus 8.7 ml/kg; $P < 0.0001$). The arterial oxygen tension/fractional inspired oxygen ratio, lung injury score, and chest radiograph scores correlated with EVLW ($r2 = 0.27$, $r2 = 0.18$, and $r2 = 0.28$, respectively; all $P < 0.0001$). **Conclusions** More than half of the patients with severe sepsis but without ARDS had increased EVLW, possibly representing subclinical lung injury. Chronic alcohol abuse was associated with increased EVLW, whereas lower EVLW was associated with survival. EVLW correlated moderately with the severity of lung injury but did not account for all respiratory derangements. EVLW may improve both risk stratification and management of patients with severe sepsis. http://ccforum.com/content/9/2/R74

114. **Breast implants following mastectomy in women with early-stage breast cancer: prevalence and impact on survival**. Gem M Le, Cynthia D O'Malley, Sally L Glaser, Charles F Lynch, Janet L Stanford, Theresa HM Keegan and Dee W West. Breast Cancer Res 2005, 7:R184-R193 doi:10.1186/bcr974. **Background** Few studies have examined the effect of breast implants after mastectomy on long-term survival in breast cancer patients, despite growing public health concern over potential long-term adverse health effects. **Methods** We analyzed data from the Surveillance, Epidemiology and End Results Breast Implant Surveillance Study conducted in San Francisco–Oakland, in Seattle–Puget Sound, and in Iowa. This population-based, retrospective cohort included women younger than 65 years when diagnosed with early or unstaged first primary breast cancer between 1983 and 1989, treated with mastectomy. The women were followed for a median of 12.4 years (n = 4968). Breast implant usage was validated by medical record review. Cox proportional hazards models were used to estimate hazard rate ratios for survival time until death due to breast cancer or other causes for women with and without breast implants, adjusted for relevant patient and tumor characteristics. **Results** Twenty percent of cases received postmastectomy breast implants, with silicone gel-filled implants comprising the most common type. Patients with implants were younger and more likely to have in situ disease than patients not receiving implants. Risks of breast cancer mortality (hazard ratio, 0.54; 95% confidence interval, 0.43–0.67) and nonbreast cancer mortality (hazard ratio, 0.59; 95% confidence interval, 0.41–0.85) were lower in patients with implants than in those patients without implants, following adjustment for age and year of diagnosis, race/ethnicity, stage, tumor grade, histology, and radiation therapy. Implant type did not appear to influence long-term survival. **Conclusions** In a large, population-representative sample, breast implants following mastectomy do not appear to confer any survival disadvantage following early-stage breast cancer in women younger than 65 years old. http://breast-cancer-research.com/content/7/2/R184

115. Conclusion

Randomization is the use of a random device to assign patients to a treatment group or control group. When the sample size is sufficiently large, randomization prevents covariate imbalance in your experiment. Randomization is not practical if patients have a strong preference for a particular treatment and is unethical if it forces some patients to endure a harmful exposure.

116. Repeat of pop quiz #5

When the demographic profile of the patients in your treatment group differ sharply from the profile of patients in your control group, you have:

1. covariate imbalance,
2. observational data,
3. response bias,
4. spectrum bias,
5. stratified data,
6. don’t know/not sure

117. Repeat of pop quiz #6

Randomization is not practical:

1. when doctors believe that the new treatment is superior to the current standard
2. when patients have a strong preference for a particular treatment
3. when the experiment requires deliberate exposure of patients to something that is known to be harmful
4. randomization is impractical for all of the above situations
5. randomization can be applied easily in all of the above situations
6. don’t know/not sure

118. Repeat of pop quiz #7

The following approaches are credible alternatives to randomization:

1. alternating between treatment and control
2. assigning all new patients to the treatment group and choosing controls from a medical database
3. assigning treatment group on the basis of the last digit of your birthday
4. letting the doctor choose whether a patient gets into the treatment group or the control group
5. none of these approaches is as effective as randomization
6. don’t know/not sure

119. It's just what the doctor ordered: observational studies.

- Abstract: An observational study is a study where the researchers do not directly intervene, but instead let the patients and/or their doctors choose the treatment. Observational studies also arise when a group is intact at the start of the study. There are four types of observational studies: cohort studies, case-control studies, cross-sectional studies, and historical control studies. While observational studies are generally considered to be less authoritative than randomized studies, with careful selection of the control subjects, observational studies can still provide persuasive results.

120. Objectives

In this class you will learn how to:

- list the four common types of observational studies,
- distinguish between cohort and case-control studies, and
- explain the limitations of historical control studies.

121. Sources

Part of the material for this webinar comes from:

- Simon SD. Statistical Evidence in Medical Trials, What Do the Data Really Tell Us? 2006. Oxford University Press: Oxford, England.
- Stats #32b: Statistical Evidence: Apples or Oranges? Randomized studies.
  - http://www.childrens-mercy.org/stats/training/hand32b.asp

122. Pop quiz #8

Which of the following is NOT an observational design?

1. Case-control study
2. Cohort study
3. Cross-sectional study
4. Historical control trial
5. Randomized control trial
6. Don’t know/not sure

123. Pop quiz #9

Which type of study is best for evaluating rare diseases:

1. Case-control study
2. Cohort study
3. Cross-sectional study
4. Historical control trial
5. Randomized control trial
6. Don’t know/not sure

124. Pop quiz #10

The historical control design is considered a weak form of evidence except when:

1. the disease being studied is rare
2. the exposure is too risky to allow random assignment
3. the mortality/morbidity rate is close to 100%
4. there is strong evidence of covariate imbalance
5. those who don’t understand history are doomed to repeat it.
6. don’t know/not sure

125. Observational studies

- There are many situations where randomization is not ethical, practical, or possible. This includes setting with:
  - a dangerous exposure,
  - limited financial resources,
  - strong patients/physicians preferences
  - groups that already exist

126. Observational studies

Observational studies are those studies where the researcher can't/won't assign patients to treatment/control groups. There are four major flavors for observational studies:

1. cohort studies,
2. case control studies,
3. cross-sectional studies, and
4. historical controls studies.

127. Cohort studies

In a cohort study, a group of patients has a certain exposure or condition. They are compared to a group of patients without that exposure or condition. Does the exposed cohort differ from the unexposed cohort on an outcome of interest?

128. Cohort studies

Example: In a study of suicide among Swedish men in the Swedish military service conscription register (Gunnell 2005), 987,308 men registered between 1968 and 1994 were divided into nine groups on the basis of four intelligence tests. These men were also linked to a Swedish cause of death register which identified a total of 2,811 suicides among these men. For each of the four intelligence tests, men scoring lower tended to have a higher rate of suicide.

129. Cohort studies

Example: In a study of psychotic symptoms in young people, a sample of young adults aged 14–24 years were divided into a group of 320 with admitted use of cannabis and a group of 2,117 did not admit to cannabis use. Both groups were followed four years later for psychotic symptoms.

130. Cohort studies

Cohort studies are intuitively appealing and selection of a control group is usually not too difficult. You have to be wary of covariate imbalance, but do not worry about every possible covariate imbalance. You should look for large imbalances, especially for covariates which are closely related to the outcome variable.

131. Cohort study

When you are studying a very rare outcome, the sample size may have to be extremely large. As a rough rule of thumb, you need to observe 25–50 outcomes in each group in order to have a reasonable level of precision. So when a condition occurs only once in every thousand patients, a cohort study would require tens of thousands of patients.

132. Cohort study

You want to avoid ‘leaky groups’ in a cohort design. If the exposure group includes some unexposed patients and the control group includes some exposed patients, then any effect you are trying to detect will be diluted.

Examples:

- Equating caffeine consumption with coffee drinking.
- Measuring dietary consumption of individuals through family shopping data.

133. Case-control study

A case-control study selects patients on the basis of an outcome, such as development of breast cancer, and are compared to a group of patients without that outcome.

134. Case-control study

Example: In a study of asthma deaths (Anderson 2005), researchers selected 532 patients who died between 1994 and 1998 with asthma mentioned in part I of the death certificate. For each asthma death, a similar asthma admission (without death) was identified at the same hospital, with a similar admission date and a similar age..

135. Case-control study

Example: In a study of vascular dementia (Chan Carusone 2004), researchers selected 28 patients with vascular dementia who were enrolled in the Geriatric Clinic at Henderson Hospital in Hamilton, Ontario, between July 1999 and October 2001. They also selected controls from a list of all caregivers at that clinic, regardless of the diagnosis of their spouse or family member, as long as the caregiver did not have any signs of dementia or stroke. Caregivers were matched by age (within 5 years) and sex. The researchers tested both cases and controls for Chalamydia.

136. Case-control study

A case-control study is very efficient in studying rare diseases. With this design, you round up all of the limited number of cases of the disease and then find a comparable control group. By contrast, a cohort design has to round up far more exposures to ensure that a handful of them will develop the rare disease.

137. Case-control study

The case-control study is always retrospective because the outcome in a case-control study has already occurred. Retrospective studies usually have more problems with data quality because our memory is not always perfect. What is worse is that sometimes the ability to remember is sharply influenced by the outcome being studied.

138. Case-control study

In a case-control study, it is often very hard to find a good control group. You want to find controls that are identical to the cases in all aspects except for the outcome itself. What does it mean to be exactly like a lung cancer patient, except for the lung cancer?

139. Case-control study

Finally, the case-control design just does not sit well with your intuition. You are trying to find factors that cause an outcome, so you are sampling from the causes while a cohort design samples from the effects. Don’t let this bother you too much, though. The mathematics that justify the case-control design were developed half a century ago (Cornfield 1951).

140. Case-control design

The careful use of the case-control design has helped answer important clinical questions which could not have been answered by other research designs. Case-control designs, for example, established the use of aspirin as a cause of Reye's syndrome (Monto 1999). It is hard to imagine how a randomized trial for Reye's syndrome could have been done.

141. Cross-sectional design

In contrast to the cohort and the case-control design, the cross-sectional study select on the basis of neither exposure nor outcome. With the cross-sectional design, you select a single group of patients and simultaneously assess both their exposure variables and their outcome variables. Typically, there are multiple exposures and multiple outcomes in a cross-sectional study.

142. Cross-sectional study

Example: In a study of intimate partner violence (Malcoe 2004), 312 Native American women attending a tribally operated clinic filled out a survey form. The survey included a modified Conflict Tactics Scale to assess whether the women experienced verbal or psychological aggression, or physical or sexual assault. The survey also asked about educational attainment, employment status, receipt of food stamps, and other questions to help determine their socioeconomic status. Since both the outcome (intimate partner violence) and the exposure (socioeconomic status) were determined at the same time, this represents a cross-sectional survey.

143. Cross-sectional study

Example: In a study of respiratory problems (Salo 2004), 5,051 seventh grade students in Wuhan, China, completed a self-administered questionnaire.These students were classified according to six respiratory outcomes (wheezing with colds, wheezing without colds, bringing up phlegm with colds, bringing up phlegm without colds, coughing with colds, coughing without colds) and two exposure variables (coal burning for cooking and cleaning, and smoking in the home). Students were not randomly assigned to an exposure; so this is an observational study. Both the outcome variables and the exposure variables were assessed at a single point in time, so this represents a cross-sectional study.

144. Cross-sectional study

Since there is no separation in time between assessment of exposure and assessment of outcome, you often cannot determine which came first. This loss of temporality makes it difficult to infer a cause-and-effect.

145. Cross-sectional study

A hypothetical example of patient height (Mann 2003), describes how a cross-sectional study might notice a negative association between height and age. Could this be because people shrink as they age, or perhaps successive generations of people are taller because of the improvements in nutrition, or perhaps taller people just die earlier? With a cross-sectional study, you cannot easily disentangle these alternate explanations.

146. Cross-sectional study

Cross-sectional studies are fast as you do not have to wait around to see what happens to the patients. These studies also allow you to easily explore relationships between multiple exposure variables and/or multiple outcome variables. But unlike the cohort design, which is useful for rare exposures, or the case-control design, which is useful for rare outcomes, the cross-sectional study is only effective if both the exposure and the outcome are relatively common events.

147. Historical controls study

In a historical controls study, researchers will assign all of the research subjects to the new therapy. The outcomes of these subjects are compared to historical records representing the standard therapy.

148. Historical controls study

Example: In a study of the rapid parathyroid hormone test (Johnson 2001), 49 patients undergoing parathyroidectomy received the rapid test. These patients were compared to 55 patients undergoing the same procedure before the rapid test was available. This is an observational study because the calendar, not the researchers, determined which test was applied. This particular observational study is a historical controls design because the control group represents patients tested before the availability of the rapid test.

149. Historical controls study

The very nature of a historical controls study guarantees that there will be a major covariate imbalance between the two groups. Thus, you have to consider any factors that have changed over time that might be related to the outcome. To what extent might these factors affect the outcome differentially?

150. Historical controls study

For the most part, historical controls are considered one of the weakest forms of evidence. The one exception is when a disease has close to 100% mortality. In that situation, there is no need for a concurrent control group, since any therapy that is remotely effective can readily be detected. Even in this situation, you want to be sure there is a biological basis for the treatment and that the disease group is homogeneous.

151. Practice exercises

- For each of the following abstracts, categorize the research studies as one of the following:
  - case-control study
  - cohort study
  - cross-sectional study
  - historical control study

152. **Body fatness during childhood and adolescence and incidence of breast cancer in premenopausal women: a prospective cohort study**. Heather J Baer, Graham A Colditz, Bernard Rosner, Karin B Michels, Janet W Rich-Edwards, David J Hunter and Walter C Willett. Breast Cancer Research 2005, 7:R314-R325 doi:10.1186/bcr998. **Introduction** Body mass index (BMI) during adulthood is inversely related to the incidence of premenopausal breast cancer, but the role of body fatness earlier in life is less clear. We examined prospectively the relation between body fatness during childhood and adolescence and the incidence of breast cancer in premenopausal women. **Methods** Participants were 109,267 premenopausal women in the Nurses' Health Study II who recalled their body fatness at ages 5, 10 and 20 years using a validated 9-level figure drawing. Over 12 years of follow up, 1318 incident cases of breast cancer were identified. Cox proportional hazards regression was used to compute relative risks (RRs) and 95% confidence intervals (CIs) for body fatness at each age and for average childhood (ages 5–10 years) and adolescent (ages 10–20 years) fatness. **Results** Body fatness at each age was inversely associated with premenopausal breast cancer incidence; the multivariate RRs were 0.48 (95% CI 0.35–0.55) and 0.57 (95% CI 0.39–0.83) for the most overweight compared with the most lean in childhood and adolescence, respectively (P for trend < 0.0001). The association for childhood body fatness was only slightly attenuated after adjustment for later BMI, with a multivariate RR of 0.52 (95% CI 0.38–0.71) for the most overweight compared with the most lean (P for trend = 0.001). Adjustment for menstrual cycle characteristics had little impact on the association. **Conclusion** Greater body fatness during childhood and adolescence is associated with reduced incidence of premenopausal breast cancer, independent of adult BMI and menstrual cycle characteristics. http://breast-cancer-research.com/content/7/3/R314

153. **Impact of a nurses' protocol-directed weaning procedure on outcomes in patients undergoing mechanical ventilation for longer than 48 hours: a prospective cohort study with a matched historical control group**. Jean-Marie Tonnelier, Gwenaël Prat, Grégoire Le Gal, Christophe Gut-Gobert, Anne Renault, Jean-Michel Boles and Erwan L'Her. Critical Care 2005, 9:R83-R89 doi:10.1186/cc3030. **Introduction** The aim of the study was to determine whether the use of a nurses' protocol-directed weaning procedure, based on the French intensive care society (SRLF) consensus recommendations, was associated with reductions in the duration of mechanical ventilation and intensive care unit (ICU) length of stay in patients requiring more than 48 hours of mechanical ventilation. **Methods** This prospective study was conducted in a university hospital ICU from January 2002 through to February 2003. A total of 104 patients who had been ventilated for more than 48 hours and were weaned from mechanical ventilation using a nurses' protocol-directed procedure (cases) were compared with a 1:1 matched historical control group who underwent conventional physician-directed weaning (between 1999 and 2001). Duration of ventilation and length of ICU stay, rate of unsuccessful extubation and rate of ventilator-associated pneumonia were compared between cases and controls. **Results** The duration of mechanical ventilation (16.6 ± 13 days versus 22.5 ± 21 days; P = 0.02) and ICU length of stay (21.6 ± 14.3 days versus 27.6 ± 21.7 days; P = 0.02) were lower among patients who underwent the nurses' protocol-directed weaning than among control individuals. Ventilator-associated pneumonia, ventilator discontinuation failure rates and ICU mortality were similar between the two groups. **Discussion** Application of the nurses' protocol-directed weaning procedure described here is safe and promotes significant outcome benefits in patients who require more than 48 hours of mechanical ventilation.
http://ccforum.com/content/9/2/R83

154. **Extravascular lung water in patients with severe sepsis: a prospective cohort study**. Greg S Martin, Stephanie Eaton, Meredith Mealer and Marc Moss. Critical Care 2005, 9:R74-R82 doi:10.1186/cc3025. **Introduction** Few investigations have prospectively examined extravascular lung water (EVLW) in patients with severe sepsis. We sought to determine whether EVLW may contribute to lung injury in these patients by quantifying the relationship of EVLW to parameters of lung injury, to determine the effects of chronic alcohol abuse on EVLW, and to determine whether EVLW may be a useful tool in the diagnosis of acute respiratory distress syndrome (ARDS). **Methods** The present prospective cohort study was conducted in consecutive patients with severe sepsis from a medical intensive care unit in an urban university teaching hospital. In each patient, transpulmonary thermodilution was used to measure cardiovascular hemodynamics and EVLW for 7 days via an arterial catheter placed within 72 hours of meeting criteria for severe sepsis. **Results** A total of 29 patients were studied. Twenty-five of the 29 patients (86%) were mechanically ventilated, 15 of the 29 patients (52%) developed ARDS, and overall 28-day mortality was 41%. Eight out of 14 patients (57%) with non-ARDS severe sepsis had high EVLW with significantly greater hypoxemia than did those patient with low EVLW (mean arterial oxygen tension/fractional inspired oxygen ratio 230.7 ± 36.1 mmHg versus 341.2 ± 92.8 mmHg; $P < 0.001$). Four out of 15 patients with severe sepsis with ARDS maintained a low EVLW and had better 28-day survival than did ARDS patients with high EVLW (100% versus 36%; $P = 0.03$). ARDS patients with a history of chronic alcohol abuse had greater EVLW than did nonalcoholic patients (19.9 ml/kg versus 8.7 ml/kg; $P < 0.0001$). The arterial oxygen tension/fractional inspired oxygen ratio, lung injury score, and chest radiograph scores correlated with EVLW ($r2 = 0.27$, $r2 = 0.18$, and $r2 = 0.28$, respectively; all $P < 0.0001$). **Conclusions** More than half of the patients with severe sepsis but without ARDS had increased EVLW, possibly representing subclinical lung injury. Chronic alcohol abuse was associated with increased EVLW, whereas lower EVLW was associated with survival. EVLW correlated moderately with the severity of lung injury but did not account for all respiratory derangements. EVLW may improve both risk stratification and management of patients with severe sepsis. http://ccforum.com/content/9/2/R74

155. **Breast implants following mastectomy in women with early-stage breast cancer: prevalence and impact on survival**. Gem M Le, Cynthia D O'Malley, Sally L Glaser, Charles F Lynch, Janet L Stanford, Theresa HM Keegan and Dee W West. Breast Cancer Res 2005, 7:R184-R193 doi:10.1186/bcr974. **Background** Few studies have examined the effect of breast implants after mastectomy on long-term survival in breast cancer patients, despite growing public health concern over potential long-term adverse health effects. **Methods** We analyzed data from the Surveillance, Epidemiology and End Results Breast Implant Surveillance Study conducted in San Francisco–Oakland, in Seattle–Puget Sound, and in Iowa. This population-based, retrospective cohort included women younger than 65 years when diagnosed with early or unstaged first primary breast cancer between 1983 and 1989, treated with mastectomy. The women were followed for a median of 12.4 years (n = 4968). Breast implant usage was validated by medical record review. Cox proportional hazards models were used to estimate hazard rate ratios for survival time until death due to breast cancer or other causes for women with and without breast implants, adjusted for relevant patient and tumor characteristics. **Results** Twenty percent of cases received postmastectomy breast implants, with silicone gel-filled implants comprising the most common type. Patients with implants were younger and more likely to have in situ disease than patients not receiving implants. Risks of breast cancer mortality (hazard ratio, 0.54; 95% confidence interval, 0.43–0.67) and nonbreast cancer mortality (hazard ratio, 0.59; 95% confidence interval, 0.41–0.85) were lower in patients with implants than in those patients without implants, following adjustment for age and year of diagnosis, race/ethnicity, stage, tumor grade, histology, and radiation therapy. Implant type did not appear to influence long-term survival. **Conclusions** In a large, population-representative sample, breast implants following mastectomy do not appear to confer any survival disadvantage following early-stage breast cancer in women younger than 65 years old. http://breast-cancer-research.com/content/7/2/R184

156. Conclusion

Observational studies are used when randomization is not possible, practical, or ethical. Cohort designs select patients on the basis of their exposure. Case-control designs select patients on the basis of their outcome. Selecting appropriate controls in a case-control design is difficult, but this design is efficient when studying a rare disease.

157. Conclusion

Cross-sectional studies select a single group of patients and classify them by multiple exposures and multiple outcomes. Because there is not always an obvious time order in the data collection, it is easy in a cross-sectional study to confuse causes and effects. Historical control studies provide an intervention to all new patients and compare them to previous medical records. Historical control studies always have a serious covariate imbalance, but are still useful when studying a condition that has close to 100% morbidity/mortality.

158. Repeat of pop quiz #8

Which of the following is NOT an observational design?

1. Case-control study
2. Cohort study
3. Cross-sectional study
4. Historical control trial
5. Randomized control trial
6. Don’t know/not sure

159. Repeat of pop quiz #9

Which type of study is best for evaluating rare diseases:

1. Case-control study
2. Cohort study
3. Cross-sectional study
4. Historical control trial
5. Randomized control trial
6. Don’t know/not sure

160. Repeat of pop quiz #10

The historical control design is considered a weak form of evidence except when:

1. the disease being studied is rare
2. the exposure is too risky to allow random assignment
3. the mortality/morbidity rate is close to 100%
4. there is strong evidence of covariate imbalance
5. those who don’t understand history are doomed to repeat it.
6. don’t know/not sure

161. Putting it all together: Meta-analyses and systematic overviews

Abstract: This class helps you assess the quality of a systematic overview or meta-analysis. In this class you will learn how to: recognize sources of heterogeneity in meta-analysis; identify and avoid problems with publication bias; and explain the ethical concerns with failure to publish and with duplicate publication.

This material is derived mainly from Chapter 5 of **Statistical Evidence in Medical Trials**.

162. Outline

1. Pop quiz
2. Introduction and motivating example
3. Were apples combined with oranges?
4. Were some apples left on the tree?
5. Repeat of pop quiz

Note: there are also issues involving study quality (were all of the apples rotten?) and practical significance (did the pile of apples amount to more than just a hill of beans?) but we will not have time to discuss those issues today.

163. Pop quiz #11

A funnel plot is useful for assessing

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don’t know

164. Pop quiz #12

Cochran's Q and I^2 are measures of

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

165. Introduction

- When there are multiple research studies evaluating a new intervention, you need to find a way to assess the cumulative evidence of these studies. You can do this informally, but medical researchers now use a formal process, known as meta-analysis. Meta-analysis, involves the quantitative pooling of data from two or more studies.

166. Introduction

- More recently, another term, systematic overview, has come into favor. A systematic overview involves the careful review and identification of all research studies associated with a topic, but it may or may not end up pooling the results of these studies. So meta-analysis represents a subset of all the systematic overviews.

167. Motivating example

- In 1992, the British Medical Journal published a controversial meta-analysis. This study (Carlsen 1992) reviewed 61 papers published from 1938 and 1991 and showed that there was a significant decrease in sperm count and in seminal volume over this period of time. For example, a linear regression model on the pooled data provided an estimated average count of 113 million per ml in 1940 and 66 million per ml in 1990.

168. Motivating example

- Several researchers (Olsen 1995; Fisch 1996) noted heterogeneity in this meta-analysis, a mixing of apples and oranges. Studies before 1970 were dominated by studies in the United States and particularly studies in New York. Studies after 1970 included many other locations including third world countries. Thus the early studies were US apples. The later studies were international oranges. There was also substantial variation in collection methods, especially in the extent to which the subjects adhered to a minimum abstinence period.

169. Motivating example

- The original meta-analysis and the criticisms of it highlight both the greatest weakness and the greatest strength of meta-analysis. Meta-analysis is the quantitative pooling of data from studies with sometimes small and sometimes large disparities. Think of it as a multicenter trial where each center gets to use its own protocol and where some of the centers are left out.

170. Motivating example

- On the other hand, a meta-analysis lays all the cards on the table. Sitting out in the open are all the methods for selecting studies, abstracting information, and combining the findings. Meta-analysis allows objective criticism of these overt methods and even allows replication of the research.

171. Motivating example

- Contrast this to an invited editorial or commentary that provides a subjective summary of a research area. Even when the subjective summary is done well, you cannot effectively replicate the findings. Since a subjective review is a black box, the only way, it seems, to repudiate a subjective summary is to attack the messenger.

172. Were apples combined with oranges?

- Meta-analyses should not have too broad an inclusion criteria. Including too broad a range of studies can lead to problems with heterogeneity (mixing apples and oranges).

173. First example of heterogeneity

- In a meta-analysis looking at antiretroviral combination therapy (Jordan 2002), both short-term and long-term outcomes were examined. A plot of duration of trial versus the log odds ratio showed that shorter duration trials of zidovudine had substantial evidence of effect (odds ratios much smaller than 1) but that the largest duration studies had little or no evidence of effect (odds ratios very close to 1).

174. Second example of heterogeneity

- Example: In a meta-analysis looking at dust mite control measures to help asthmatic patients (Gotzsche 1998), the studies exhibited heterogeneity across several factors.

175. Second example of heterogeneity

- Type of intervention:
  - six examined chemical interventions,
  - thirteen examined physical interventions,
  - four examined a combination approach.
- Research design:
  - nine of these trials were crossovers,
  - fourteen had a parallel control group.
- Blinding
  - seven studies had no blinding,
  - three studies had partial blinding,
  - thirteen studies used a double blind.

176. Second example of heterogeneity

- Age of patients
  - nine studies the average age of the patients was only 9 or 10 years,
  - nine other studies had an average age of 30 or more,
  - five studies had a greater mix of ages.
- Duration
  - eleven studies lasted eight weeks or less,
  - five studies lasted a full year,
  - seven studies had an intermediate duration

177. Possible sources of heterogeneity

- This list is adapted from Horwitz 1987
  - Inclusion/exclusion criteria
  - Geographical limitations
  - Independent versus matched controls
  - Dose/timing of drug administration
  - Length of follow-up
  - Drop-out rates
  - Allowable physician discretion
  - Outcome measure

178. Measuring heterogeneity

- Cochran's Q: A value close to the number of studies is good, but a value much larger is bad.
- I^2: ranges between 0% and 100%, larger values indicating greater heterogeneity.
- Many researchers recommend a qualitiative assessment of heterogeneity.

179. Forest plot

- The forest plot provides a graphical summary of the studies. This plot can be used to evaluate heterogeneity.
  - Location of square represents the point estimate,
  - Size of square represents weight associated with that estimate, and
  - Lines drawn to upper and lower confidence limits.

180. Forest plot

- Look for marked departures from a normal random scatter:
  - Most studies cluster together, but one or two outlying studies (but okay if outlying studies have small sample sizes).
  - Bimodal patterns (e.g., half the studies show a strong effect, half show little or no effect).

181. Forest plot example

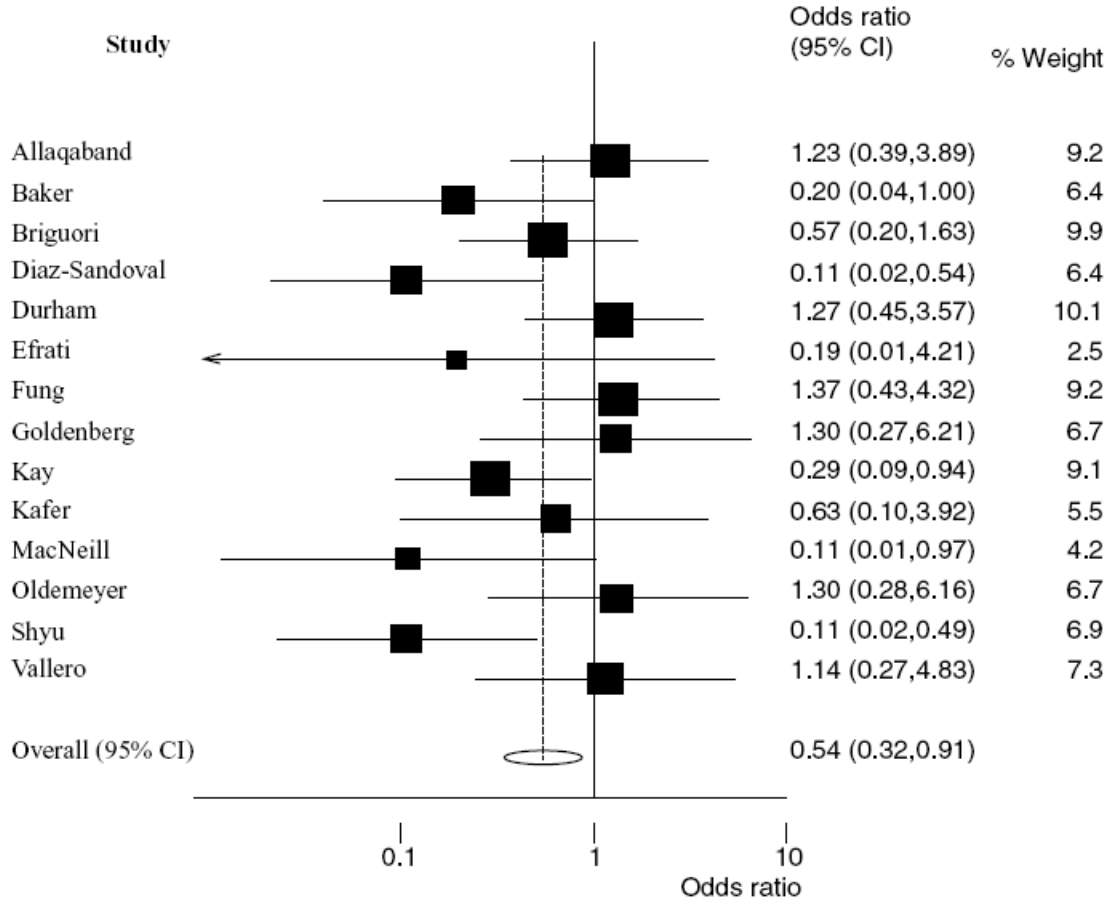


Figure 5.1 This image is from an open source publication. You can find the original article at www.biomedcentral.com/1741-7015/2/38 and this particular figure at www.biomedcentral.com/1741-7015/2/38/figure/F2.

182. Handling heterogeneity

- There are several common approaches for coping with heterogeneity
  - Strict inclusion/exclusion criteria
  - Sensitivity/subgroup analysis
  - Meta-regression
  - “Just say no”

183. Example of strict inclusion/exclusion criteria

- A meta-analysis of topical NSAIDs for musculoskelatal pain (Mason 2004) identified 60 target papers, but for 12 of the papers, there was no data that could be extracted for a meta-analysis. An additional 23 studies were removed based on the following exclusion criteria:
  - no studies for mouth or eye diseases;
  - no studies where fewer than 10 patients were randomized to the treatment;
  - no studies where treatment occurred less frequently than daily;
  - no observational studies; and
  - no unblinded studies.

184. First example of strict sensitivity/subgroup analysis

- In a study of extra corporeal shock wave therapy for plantar heel pain (Thomson 2005), six studies met the researchers inclusion criteria, but one study did not report a standard deviation for the outcome measure. The authors were forced to estimate what the standard deviation should be for this study. As a quality check, they also ran a meta-analysis without this study and found that a modest effect in favor of the therapy was no longer statistically significant.

185. Second example of strict sensitivity/subgroup analysis

- In a study of topical NSAIDs for osteoarthritis and tendinitis (Mason 2004), researchers identified 25 trials relating to efficacy or harm, including 14 placebo-controlled trials. These studies varied substantially in
  - quality scores,
  - number of patients studied,
  - type of outcome measure (physician determined versus self report) and
  - condition being treated (osteoarthritis versus other musculoskeletal conditions).
- But when the results were tabulated separately for low and high quality scores, small and large studies, etc., there were no statistically significant differences.

186. Meta-regression

- You can use meta-regression to try to adjust for heterogeneity in a metaanalysis. In meta-regression, each study becomes a data point, and various study characteristics, such as the severity of illness at baseline, the dose of the medication being given, etc. become independent variables. This is an approach that would work very similarly to the adjustment for covariates in a regression model. The result, meta-regression, is an area of active research and looks to be a promising way to handle heterogeneity in a more rigorous fashion.

187. Example of meta-regression

- In a study of diagnostic tests for endometrial hyperplasia (Clark 2004), researchers identified 27 studies using miniature endometrial biopsy devices or ultrasonography. In some of the studies, verification of the diagnosis was delayed by more than 24 hours. Although the ability to discriminate between diseased and healthy patients was present in most studies, the discriminatory power, as measured by the diagnostic odds ratio was four times weaker among studies with delayed verification than studies with no delay.

188. “Just say no”

- If the degree of heterogeneity is too extreme, you should just say no and refuse to run a meta-analysis. You can still discuss the studies in a qualitative fashion, but do not try to compute an overall estimate of effect because that estimate would be meaningless.

189. Example of “Just say no”

- In a systematic review of beta-2 agonists for treating chronic obstructive pulmonary disease (Husereau 2004), researchers identified 12 studies. But the authors could not pool the results because they
  - “found that even commonly measured outcomes, such as FEV1, could not be combined by meta-analysis because of differences in how they were reported. For example, in the six trials comparing salmeterol with placebo, FEV1 was reported as a mean change in percent predicted, a mean change overall, a mean difference between trial arms, no difference (without data), baseline and overall FEV1 (after 24 hrs without medication) and as an 0 to 12 hour area-under-the-curve (FEV1-AUC) function. We were not successful in obtaining more data from study authors. We also had concerns about the meta-analysis of data from trials of parallel and crossover design and differences in spirometry protocols including allowable medications. Therefore, we decided on a best evidence synthesis approach instead.”

190. On your own

- Read the following excerpts and comment on the degree to which heterogeneity is present among the studies being examined.

191. Psychoeducation for depression, anxiety and psychological distress: a meta-analysis (1 of 3)

- Studies were included if: the psychoeducation targeted depression, anxiety or psychological distress; participants were described as either experiencing mood or anxiety disorders; or if they experienced elevated scores (equal to or above a specified cut-off score, see Table 1) on depression, anxiety or psychological distress scales. To be included, studies were required to have a randomized controlled design, which incorporated a no intervention, attention-placebo or a waitlist control group to which psychoeducation was compared. All included studies were required to report mental health outcomes (depression, anxiety or psychological distress) and were published in peer-reviewed, English language journals. There was no restriction on the age of participants.
  - www.biomedcentral.com/1741-7015/7/79

192. Psychoeducation for depression, anxiety and psychological distress: a meta-analysis (2 of 3)

- Studies were excluded if the education component was offered in addition to other components (for example, psychotherapy with elements of psychoeducation or psychoeducation enhanced with treatment as usual) or when the intervention was compared solely to a (potentially) active treatment (for example, medication, treatment as usual or psychotherapy). Studies were also excluded: when the intervention was not passive psychoeducation but involved an active intervention (for example, components of CBT or IPT, relaxation exercises or homework or group discussion); or when psychoeducation was aimed at target groups where there was a concomitant physical health or mental disorder; or where the target of the intervention was a carer or parent of the person with anxiety or depression (for example, medical illness, other mental health disorders, parental programmes, family-caregiver programmes).

193. Psychoeducation for depression, anxiety and psychological distress: a meta-analysis (3 of 3)

- Of the five relevant papers, four papers describing three studies used depressive symptoms or disorders as primary outcome measure, while one study reported psychological distress as an outcome measure (see Table 1). Two studies used evidence-based medical/psychological depression/anxiety information; one of them also gave advice. Two studies used mailed feedback based on test results and provided advice and one study used leaflets as intervention type. Two papers reporting one study used a website. Two studies compared the intervention with an attention placebo-control, while two studies compared the intervention to no intervention condition. One study recruited participants from the community, one study used primary care participants, one study recruited employees and one study included college students. A total of 694 participants were recruited across all the studies. All included studies used individual rather than group formats. Interventions across all studies ranged from one single email or leaflet to six sessions of psychoeducation.

194. Traditional Chinese medicines in the treatment of hepatocellular cancers: a systematic review and meta-analysis (1 of 2)

- To be eligible for inclusion in our systematic review, studies had to have enrolled adult patients (>18 years) with liver cancer. The patients had to be randomly allocated to an active TCM formulation treatment or a control group with either placebo or no treatment. In addition, any co-intervention had to be the same in both groups except for the TCM formulation. We excluded studies that reported only laboratory values rather than clinical responses. We also excluded direct comparisons of TCM formulations
- **TCM Interventions**: The TCM interventions identified in this study were principally combinations of different herbal medicines or animal/insect extracts (Additional file 1). A brief outline on the oncologic and immunologic pharmacology of the most commonly used ingredients is presented below.
- **Astragalus**: Astragalus appears to have a number of immunomodulatory properties [55-57]. Astragalus appears to have anti-tumour activity where its potentiates LAK cell activity in vitro when used in combination with IL-2[58]. Astragalus appears to restore in vitro T-cell function, which is suppressed in cancer patients[59].
  - www.jeccr.com/content/28/1/112

195. Traditional Chinese medicines in the treatment of hepatocellular cancers: a systematic review and meta-analysis (2 of 2)

- **Panax ginseng**: Panax ginseng and its chemical constituents were found to have inhibitory effects on putative carcinogenesis mechanisms, e.g., cell proliferation and apoptosis, immunosurveillance and angiogenesis[60]. Ginsenosides from Panax ginseng have been shown to inhibit tumor cell invasion and to suppress sister chromatid exchanges in human lymphocytes[61].
- **Toad skin secretions (bufotoxin):** The toad skin secretion bufalin was found to induce apoptosis in human-leukemia cells by altering expression of apoptotic genes c-myc and bcl-2[62]. Other toad skin secretions like 3-formyloxyresibufogenin, 19-oxobufalin, 19-oxodesacetylcinobufagin, 6-hydroxycinobufagin and 1-hydroxybufalin were found to exert inhibitory effects on KB, HL-60 and MH-60 cancer cell lines[63].
- **Beetle extracts (Mylabris)**: An extract from Mylabris phaleratais, the dried body of the Chinese blister beetle, was shown to have anti-cancer activity via inducing cancer cell apoptosis and was associated with little toxicity[64].
- **Atractylodes**: Atractylodes appears to have anticancer activity by inducing apoptosis and cytotoxic effects against leukemia and other cancer cell lines[65].
- **Bupleurum**: Saikosaponins from Bupleurum falcatum were shown to exhibit potent anti-cell adhesive activity on solid tumour cells and to have strong hemolytic action[66].
- **Curcuma**: Curcuma longa may have immunostimulatory activity[67].

196. Were some apples left on the tree?

- Publication bias: the tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings. There is solid empirical evidence (e.g., Dickersin 1990) that negative studies are less likely to be published.

197. Ethical concerns with failure to publish

- Researchers who fail to publish their research, however, are behaving unethically (Chalmers 1990). These research studies almost always use human volunteers. These volunteers might be participating because they need the money or perhaps they are curious about the scientific process. But many of them volunteer because they want to help others who have the same disease or condition. These volunteers submit themselves willingly to some level of inconvenience, and possibly additional pain and risk. If you ask these volunteers to make this sacrifice, but you do not publish, you have abused their good will.

198. Should unpublished studies be included?

- The inclusion of unpublished studies, however, is controversial. At least one reference (Cook 1993), has argued that unpublished studies have failed to meet a basic quality screen, the peer review process. Including studies that have not been peer reviewed will lower the overall quality of the meta-analysis. This opinion, however, is in the minority, and most experts in meta-analysis suggest that you include unpublished studies if you can find them. Failure to include unpublished studies can lead to serious bias.

199. Duplicate publication

- Duplicate publication is the flip side of the same coin. The data from some studies may appear twice (or even three times) in the peer-reviewed literature, without appropriate attribution. If you double count these studies accidentally, you will produce a biased result because duplicate publications are more likely to be positive.

200. Ethical concerns with duplicate publication

- Duplicate publication raises serious ethical issues:
  - Violation of copyright
  - Padding of resumes
  - Abuse of volunteer services of referees/editors
  - Taking page space away from other deserving publications.
- There are reasonable justifications for duplicate publication, such as translating a publication into English to insure a wider dissemination of the research findings. These exceptions, however, would always have an obvious citation of the original source.

201. Example of duplicate publication

- In 84 studies of the effect of ondansetron on postoperative emesis, 14 (17%) were second or even third time publications of the same data-set (Tramer 1997). The duplicate studies had much larger effects and adding the duplicates to the originals produced an overestimation of treatment efficacy of 23%. Tracking down the duplicate publications was quite difficult. More than 90% of the duplicate publications did not crossreference the other studies. Four pairs of identical trials were published by completely different authors without any common authorship.

202. Don’t rely exclusively on Medline

- While a Medline search is a very effective way to identify published research, it should not be the only source of publications for a meta-analysis. There are many important journals which are not included in Medline. It is hard to get an accurate count of how many journals do NOT appear in Medline, but the numbers appear to be substantial. You might suspect that journals indexed by Medline are more prestigious and more likely to publish positive findings than other journals, but I am unaware of any data to substantiate this. Still, a search that included only Medline articles would be considered grossly inadequate in most situations.

203. Don’t rely English-language only publications

- Some meta-analyses restrict their attention to English language publications only. While this may seem like a convenience, in some situations, researchers might tend to publish in an English language journal for those trials which are positive, and publish in a (presumably less prestigious) native language journal for those trials which are negative (Gregoire 1995). Restrictions to English language only publications is especially troublesome for complementary and alternative medicine, since so much of this research appears in non-English language journals.

204. Using a funnel plot to detect publication bias

- The most common approach to evaluate publication bias is to use a funnel plot. The funnel plot displays
  - the results of the individual studies (e.g. the log odds ratio) on the horizontal axis,
  - the size of the study (or sometimes the standard error of the study) on the vertical axis.
- Often a reference line is drawn at the value that represents no effect.

205. Using a funnel plot to detect publication bias

- The rationale behind this plot
  - big studies get published no matter what the result
  - smaller studies are subject to publication bias
- If there is no publication bias, then the funnel plot should show symmetry for both small sample sizes and large sample sizes, though you should expect to see less variation as the sample size increases. This leads to a funnel shape.

206. Example of a funnel plot

- The rationale behind this plot
  - big studies get published no matter what
  - smaller studies are subject to publication bias
- If there is no publication bias, then the funnel plot should show symmetry for both small sample sizes and large sample sizes, though you should expect to see less variation as the sample size increases. This leads to a funnel shape.
- Although funnel plots are commonly used, there is some suggestion that they are not effective.

207. Funnel plot example showing symmetry

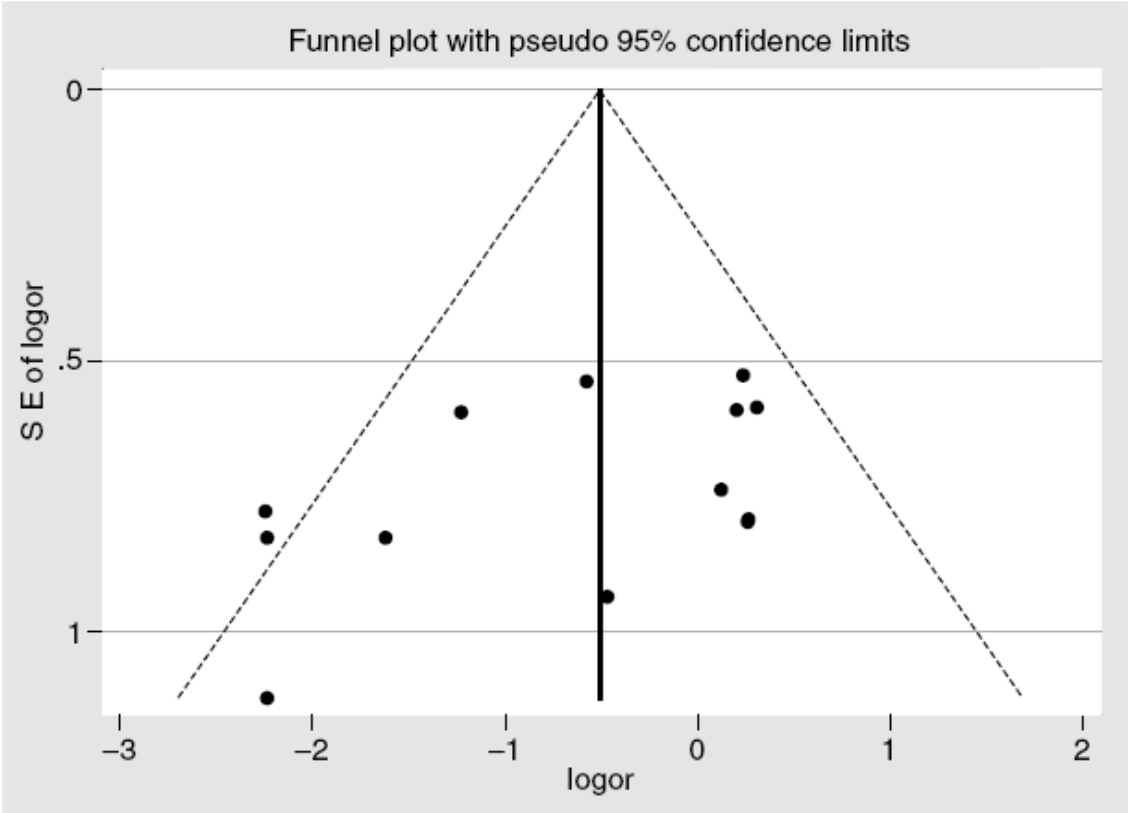


Figure 5.3 This image is from an open source publication. You can find the original article at www.biomedcentral.com/1741-7015/2/38 and this particular figure at www.biomedcentral.com/1741-7015/2/38/figure/F4.

208. Funnel plot example showing possible publication bias

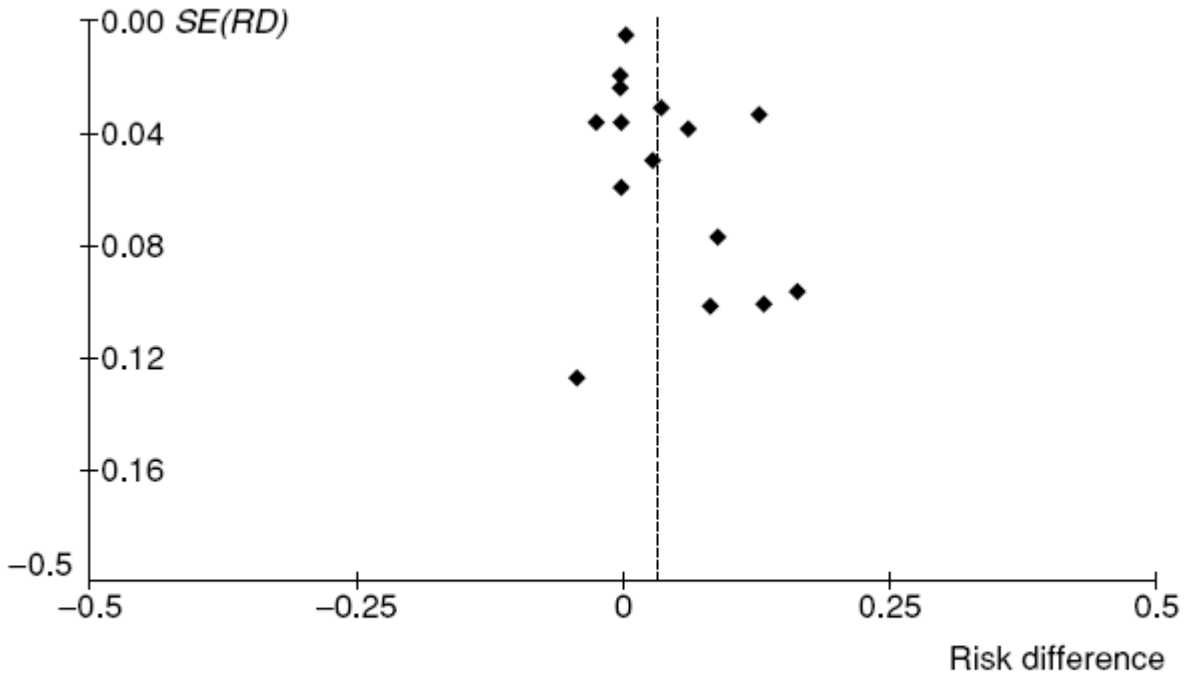


Figure 5.4 This image is from an open source publication. You can find the original article at www.biomedcentral.com/1741-7015/2/11 and this particular figure at www.biomedcentral.com/1741-7015/2/11/figure/F3.

209. How to avoid or minimize problems with publication bias

1. Use several bibliographic databases, not just Medline.
2. Search through registries of clinical trials.
3. Hand search through specialized journals
4. Examine bibliographies of articles found on first pass through.
5. Examine “gray literature” (presentations, dissertations, etc.)
6. Send out letter to prominent leaders in the area asking for help.

210. On your own

- Read the following excerpts and comment on the extent to which the researchers went to avoid publication bias.

211. Balloon kyphoplasty in malignant spinal fractures: a systematic review and meta-analysis

- A systematic literature search was carried out up to September 2008 using several databases (MEDLINE, EMBASE, CINAHL, ISI Proceedings, The Cochrane Library, DARE, NHS EED and the HTA Database of the CRD). The search strategy was: #1: (balloon kyphoplasty), #2: (fracture*) or (vertebra*) or (neoplasm*) or (tumor*), #3: #1 and #2. There were no language restrictions. The search was completed manually using references from identified studies and reviews [17], and contact was made with experts in the field. No contact was made with industry.
  - www.biomedcentral.com/1472-684X/8/12

212. Efficacy of pharmacotherapies for short-term smoking abstinance: A systematic review and meta-analysis

- In consultation with a medical librarian (PR), we established a search strategy. We searched independently, in duplicate, the following 10 databases (from inception to October 1, 2008): MEDLINE, EMBASE, Cochrane CENTRAL, AMED, CINAHL, TOXNET, Development and Reproductive Toxicology, Hazardous Substances Databank, Psych-info and Web of Science, databases that included the full text of journals (OVID, ScienceDirect, and Ingenta, including articles in full text from approximately 1700 journals since 1993). In addition, we searched the bibliographies of published systematic reviews.[5,19,7,10,11,13,26] and health technology assessments.[27] Searches were not limited by language, sex or age.
  - www.harmreductionjournal.com/content/6/1/25

213. Repeat of pop quiz #11

A funnel plot is useful for assessing

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don’t know

214. Repeat of pop quiz #12

Cochran's Q and I^2 are measures of

1. heterogeneity
2. publication bias
3. study quality
4. not sure/don't know

215. Conclusion

Where do you go from here?

1. Don’t pretend that you are a professional statistician, no matter how well I taught this course.
2. But, you should be a much better consumer of Statistics.
3. You are in a better position to raise questions that your customers need to ask when they read a paper.

Sign up for The Monthly Mean, www.pmean.com/news