Supplemental Material

CBE—Life Sciences Education

Masel *et al.*

Supplementary Materials for "*Evidence-based medicine as a tool for undergraduate probability and statistics education*"

Table of Contents

Table S1. Pre-course, post-course, and changes in QRQ total sub-scores for 2015 (hybrid).

Category	Category text	pre-test mean[1]	pre-test stdev	post-test mean[1]	post-test stdev	mean difference	*P*
Overall	-	61.88	12.70	65.45	12.23	3.57 (±3.46)	0.046
Competencies	*Correctly interprets probabilities*	2.00	1.28	2.36	1.45	0.36 (±0.65)	0.283
	Correctly interprets measures of central tendency	3.92	0.75	4.00	0.80	0.08 (±0.27)	0.547
	Understands how to select an appropriate average	2.48	0.91	2.33	0.89	-0.14 (±0.37)	0.449
	Correctly computes probability	2.75	1.43	2.68	1.25	-0.07 (±0.63)	0.821
	Understands independence	4.26	0.80	4.43	0.57	0.17 (±0.32)	0.305
	Understands sampling variability	2.71	0.66	2.90	0.92	0.18 (±0.36)	0.320
	Distinguishes between correlation and causation	3.76	1.35	4.29	0.85	0.52 (±0.42)	0.019
	Correctly interprets two-way tables	3.36	1.81	3.07	2.00	-0.29 (±0.67)	0.404
	Understands importance of large samples	2.93	1.59	3.50	1.40	0.57 (±0.71)	0.118
	Understands sources of bias and error	4.26	0.84	4.46	0.62	0.20 (±0.28)	0.165
	Recognizes features of good experimental design	3.54	1.17	3.68	1.22	0.14 (±0.51)	0.581
Misconceptions	*Misconceptions involving averages*	2.41	0.72	2.27	0.72	-0.14 (±0.24)	0.244
	Outcome orientation misconception	1.49	0.39	1.43	0.36	-0.06 (±0.13)	0.363
	Good samples have to represent a high percentage of the population	2.82	1.33	2.54	1.45	-0.29 (±0.58)	0.334
	Law of small numbers	2.21	1.37	1.93	1.02	-0.29 (±0.53)	0.293
	Representativeness misconception	1.62	0.63	1.50	0.53	-0.12 (±0.29)	0.421
	Correlation implies causation	2.24	1.35	1.71	0.85	-0.52 (±0.42)	0.019
	Equiprobability bias	3.25	1.38	3.54	1.40	0.29 (±0.62)	0.362
	Groups can only be compared if they are the same size	2.29	1.90	2.14	1.84	-0.14 (±0.87)	0.745
	Failure to distinguish the difference between a sample and a population	2.21	0.99	2.14	1.01	-0.07 (±0.38)	0.713
	Failure to consider and evaluate all of the data	1.07	0.26	1.11	0.31	0.04 (±0.16)	0.663
	Inability to create and evaluate fractions or percents	1.43	0.40	1.37	0.39	-0.06 (±0.19)	0.523
	Only large effects can be considered meaningful	1.86	1.67	2.43	1.95	0.57 (±0.79)	0.161
	Failure to recognize potential sources of bias and error	1.69	0.74	1.50	0.53	-0.19 (±0.25)	0.133
	Assumes more decimal places indicate greater accuracy	1.14	0.76	1.00	0.00	-0.14 (±0.29)	0.326
	Inability to interpret probabilities	1.25	0.22	1.23	0.18	-0.02 (±0.09)	0.646

[1] Overall scores scaled to 0–100%; individual scores scaled to 1–5 Likert-like scale. Error on mean differences is ± 2 x s.e.m. of paired post vs. pre-test differences.

Table S2. Pre-course, post-course, and changes in ATS total sub-scores for 2015 (hybrid).

Category	pre-test mean[1]	pre-test stdev	post-test mean[1]	post-test stdev	mean difference	*P*
Overall Attitude	69.41	12.40	72.96	11.20	3.54 (±3.85)	0.077
I feel that statistics will be useful to me in my profession	4.07	0.73	4.19	0.83	0.11 (±0.36)	0.536
The thought of being inrolled in a statistics course makes me nervous	2.93	1.49	3.04	0.90	0.11 (±0.48)	0.594
A good researcher must have training in statistics	4.48	0.58	4.70	0.47	0.22 (±0.29)	0.144
Statistics seems very mysterious to me	3.07	1.11	3.85	0.72	0.78 (±0.37)	0.001
Most people would benefit from taking a statistics course	4.19	0.79	4.41	0.57	0.22 (±0.25)	0.095
I have difficulty seeing how statistics relates to my field of study	3.96	0.98	4.30	0.54	0.33 (±0.37)	0.090
I see being enrolled in a statistics course as a very unpleasant experience	3.48	1.16	3.37	0.93	-0.11 (±0.40)	0.603
I would like to continue my statistical training in an advanced course	2.89	0.75	2.89	0.97	0 (±0.40)	0.957
Statistics will be useful to me in comparing the relative merits of different objects, methods, programs, etc.	4.11	0.58	4.22	0.58	0.11 (±0.33)	0.524
Statistics is not really very useful because it tells us what we already know anyway	4.22	0.64	4.56	0.51	0.33 (±0.21)	0.008
Statistical training is relevant to my performance in my field of study	3.89	0.70	4.00	0.78	0.11 (±0.34)	0.539
I wish that I could have avoided taking my statistics course	3.56	0.89	3.78	0.75	0.22 (±0.33)	0.198
Statistics is a worthwhile part of my professional training	3.89	0.70	3.85	0.66	-0.04 (±0.29)	0.830
Statistics is too math oriented to be of much use to me in the future	4.04	0.71	3.96	0.59	-0.07 (±0.28)	0.618
I get upset at the thought of enrolling in another statistics course	3.22	1.05	3.26	0.90	0.04 (±0.43)	0.891
Statistical analysis is best left to the "experts" and should not be part of a lay professional's job	3.96	0.59	4.11	0.70	0.15 (±0.30)	0.358
Statistics is an inseparable aspect of scientific research	4.15	0.66	4.48	0.51	0.33 (±0.24)	0.015
I feel intimidated when I have to deal with mathematical formulas	3.30	1.17	3.26	0.90	-0.04 (±0.38)	0.884
I am excited at the prospect of actually using statistics in my job	3.33	0.83	3.19	0.83	-0.15 (±0.33)	0.378
Studying statistics is a waste of time	4.33	0.55	4.33	0.62	0 (±0.24)	1.000
My statistical training will help me better understand the research being done in my field of study	4.11	0.58	4.22	0.64	0.11 (±0.22)	0.351
One becomes a more effective "consumer" of research findings if one has some training in statistics	4.19	0.56	4.41	0.57	0.22 (±0.31)	0.167
Training in statistics makes for a more well-rounded professional experience	4.15	0.36	4.37	0.49	0.22 (±0.22)	0.066
Statistical thinking can play a useful role in everyday life	3.96	0.52	4.26	0.66	0.3 (±0.23)	0.024
Dealing with numbers makes me uneasy	3.67	1.07	3.56	0.89	-0.11 (±0.34)	0.539
I feel that statistics should be required early in ones professional training	3.26	0.6	3.63	0.79	0.37 (± 0.27)	0.003
Statistics is too complicated for me to use effectively	3.63	0.82	3.74	1	0.11 (± 0.32)	0.830
Statistical training is not really useful for most professionals	3.68	0.77	3.89	0.8	0.21 (± 0.26)	0.850
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write	2.68	0.84	2.53	0.89	-0.16 (± 0.3)	0.321

[1] Overall scores scaled to 0–100%; individual scores reflect 1–5 Likert scale. Scores for all items oriented to reflect 1–5 negative-to-positive transition, with 3 being neutral. Error on mean differences is ± 2 x s.e.m. of paired post vs. pre-test differences. P-values were obtained from non-parametric paired Wilcoxon signed-rank tests.

Table S3. Pre-course, post-course, and changes in QRQ total sub-scores for 2014 and 2015 combined.

Category	Category text	pre-test mean[1]	pre-test stdev	post-test mean[1]	post-test stdev	mean difference	*P*
Overall	-	60.15	11.14	65.27	12.17	5.12 (±2.24)	<0.001
Competencies	*Correctly interprets probabilities*	2.32	1.37	2.65	1.46	0.32 (±0.37)	0.086
	Correctly interprets measures of central tendency	3.82	0.77	3.93	0.77	0.12 (±0.18)	0.204
	Understands how to select an appropriate average	2.43	0.90	2.53	0.96	0.10 (±0.26)	0.450
	Correctly computes probability	2.76	1.24	2.56	1.18	-0.21 (±0.39)	0.295
	Understands independence	4.01	1.08	4.31	0.86	0.30 (±0.29)	0.038
	Understands sampling variability	2.62	0.74	3.15	0.89	0.53 (±0.23)	0.000
	Distinguishes between correlation and causation	3.65	1.27	4.14	1.15	0.49 (±0.28)	0.001
	Correctly interprets two-way tables	3.21	1.80	3.15	1.87	-0.06 (±0.56)	0.833
	Understands importance of large samples	3.06	1.62	3.71	1.46	0.65 (±0.49)	0.011
	Understands sources of bias and error	4.07	0.86	4.34	0.76	0.27 (±0.21)	0.014
	Recognizes features of good experimental design	3.65	1.13	3.62	1.22	-0.03 (±0.32)	0.854
Misconceptions	*Misconceptions involving averages*	2.49	0.71	2.23	0.68	-0.26 (±0.19)	0.008
	Outcome orientation misconception	1.51	0.38	1.41	0.35	-0.11 (±0.10)	0.030
	Good samples have to represent a high percentage of the population	2.63	1.30	2.47	1.38	-0.16 (±0.36)	0.378
	Law of small numbers	2.12	1.26	1.74	0.97	-0.38 (±0.36)	0.036
	Representativeness misconception	1.80	0.80	1.56	0.68	-0.25 (±0.21)	0.021
	Correlation implies causation	2.35	1.27	1.86	1.15	-0.49 (±0.28)	0.001
	Equiprobability bias	3.25	1.38	3.47	1.35	0.22 (±0.43)	0.310
	Groups can only be compared if they are the same size	2.18	1.84	2.35	1.91	0.18 (±0.57)	0.536
	Failure to distinguish the difference between a sample and a population	2.09	1.00	1.91	1.00	-0.18 (±0.27)	0.203
	Failure to consider and evaluate all of the data	1.18	0.42	1.18	0.46	0.00 (±0.13)	1.000
	Inability to create and evaluate fractions or percents	1.41	0.41	1.45	0.46	0.03 (±0.14)	0.621
	Only large effects can be considered meaningful	1.94	1.71	2.00	1.74	0.06 (±0.52)	0.820
	Failure to recognize potential sources of bias and error	1.85	0.79	1.58	0.65	-0.27 (±0.19)	0.005
	Assumes more decimal places indicate greater accuracy	1.12	0.68	1.12	0.68	0.00 (±0.24)	1.000
	Inability to interpret probabilities	1.28	0.25	1.25	0.22	-0.03 (±0.07)	0.349

[1] Overall scores scaled to 0–100%; individual scores scaled to 1–5 Likert-like scale. Error on mean differences is ± 2 x s.e.m. of paired post vs. pre-test differences.

Table S4. Pre-course, post-course, and changes in ATS total sub-scores for 2014 and 2015 combined.

Category	pre-test mean[1]	pre-test stdev	post-test mean[1]	post-test stdev	mean difference	*P*
Overall Attitude	66.04	11.49	70.91	12.06	4.87 (±2.66)	0.001
I feel that statistics will be useful to me in my profession	3.94	0.86	4.18	0.91	0.24 (±0.25)	0.045
The thought of being inrolled in a statistics course makes me nervous	2.97	1.21	3.06	1.02	0.09 (±0.28)	0.506
A good researcher must have training in statistics	4.17	0.80	4.62	0.52	0.45 (±0.22)	0.000
Statistics seems very mysterious to me	3.21	1.00	3.67	0.81	0.45 (±0.26)	0.001
Most people would benefit from taking a statistics course	3.86	0.78	4.17	0.67	0.30 (±0.17)	0.001
I have difficulty seeing how statistics relates to my field of study	3.86	0.94	4.15	0.85	0.29 (±0.24)	0.017
I see being enrolled in a statistics course as a very unpleasant experience	3.23	1.06	3.48	0.92	0.25 (±0.29)	0.093
I would like to continue my statistical training in an advanced course	2.70	0.74	2.91	1.03	0.21 (±0.25)	0.109
Statistics will be useful to me in comparing the relative merits of different objects, methods, programs, etc.	3.94	0.60	4.06	0.72	0.12 (±0.22)	0.289
Statistics is not really very useful because it tells us what we already know anyway	4.11	0.66	4.41	0.63	0.30 (±0.17)	0.001
Statistical training is relevant to my performance in my field of study	3.75	0.81	4.03	0.79	0.28 (±0.20)	0.009
I wish that I could have avoided taking my statistics course	3.30	1.02	3.67	1.00	0.36 (±0.26)	0.010
Statistics is a worthwhile part of my professional training	3.75	0.71	3.89	0.73	0.14 (±0.19)	0.151
Statistics is too math oriented to be of much use to me in the future	4.00	0.72	4.05	0.81	0.05 (±0.20)	0.664
I get upset at the thought of enrolling in another statistics course	3.14	1.03	3.29	1.06	0.15 (±0.30)	0.286
Statistical analysis is best left to the "experts" and should not be part of a lay professional's job	3.85	0.75	3.94	0.70	0.09 (±0.21)	0.386
Statistics is an inseparable aspect of scientific research	4.08	0.74	4.25	0.79	0.17 (±0.18)	0.078
I feel intimidated when I have to deal with mathematical formulas	3.20	1.09	3.35	1.05	0.15 (±0.25)	0.248
I am excited at the prospect of actually using statistics in my job	3.00	0.89	3.16	0.95	0.16 (±0.28)	0.331
Studying statistics is a waste of time	4.23	0.63	4.28	0.78	0.05 (±0.19)	0.674
My statistical training will help me better understand the research being done in my field of study	4.00	0.71	4.12	0.76	0.12 (±0.15)	0.107
One becomes a more effective "consumer" of research findings if one has some training in statistics	3.98	0.60	4.22	0.74	0.23 (±0.23)	0.055
Training in statistics makes for a more well-rounded professional experience	4.02	0.41	4.18	0.68	0.17 (±0.18)	0.078
Statistical thinking can play a useful role in everyday life	3.80	0.59	4.06	0.79	0.26 (±0.19)	0.008
Dealing with numbers makes me uneasy	3.62	1.03	3.66	0.99	0.05 (±0.21)	0.657
I feel that statistics should be required early in ones professional training	3.43	0.64	3.88	0.78	0.45 (±0.21)	0.000
Statistics is too complicated for me to use effectively	3.68	0.79	3.72	0.89	0.05 (±0.23)	0.770
Statistical training is not really useful for most professionals	3.83	0.72	3.94	0.77	0.11 (±0.20)	0.286
Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write	2.78	0.89	2.78	1.07	0.00 (±0.27)	0.952

[1] Overall scores scaled to 0–100%; individual scores reflect 1–5 Likert scale. Scores for all items oriented to reflect 1–5 negative-to-positive transition, with 3 being neutral. Error on mean differences is ± 2 x s.e.m. of paired post vs. pre-test differences. P-values were obtained from non-parametric paired Wilcoxon signed-rank tests.

Ecology 379: Evidence Based Medicine **Name:**
Spring 2014
Final Exam

1(a). (4 points) For each statement below about a criminal trial, mark an "F" for those that describe a "forward probability" and "B" for those that describe a "backward probability" of the sort you need to use Bayes' theorem to study.

i. The probability that a man is innocent given that goes to jail
ii. The probability that a man goes to jail given that he is innocent
iii. The probability that a man is guilty given that he goes to jail
iv. The probability that a man goes to jail given that he is guilty

1(b). (4 points) Which pair or pairs of probabilities above must add up to 1?

1(c). (6 points) Assuming that the court system works on the basis of "innocent until proven guilty", describe, using similar language to part (a), the probability of making a type I error and the probability of making a type II error.

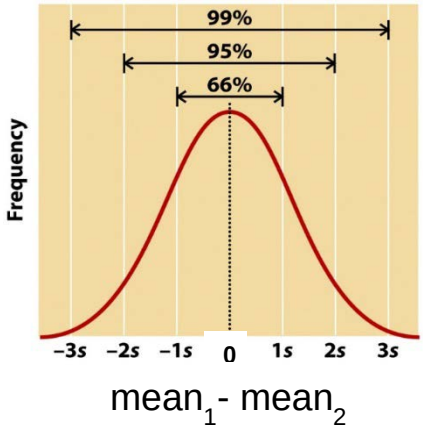
2(a). (6 points) Adderall is used by many students because they believe it helps them study and do well on exams. In a class of 20 students, 3 are taking Adderall for this reason, while the other 17 do not take Adderall. The best, second-best and third-best grade all go to the 3 students taking Adderall. If Adderall does NOT affect exam performance, what is the probability of this happening?

2(b). (5 points) You want to know whether students should take Adderall when they wish do to well in the class, and they are willing to risk the side effects. What are the weaknesses of the experimental design described in part (a) for answering this question? What sort of experiment would be better at answering this question?

3(a). (5 points) There are 100 students enrolled in a science class. The average score on the common final exam is 80%, with a standard deviation of 10 percentage points. What is the standard error associated with the 80% mean final score? (*You will need to use your answer to part (a) at least once in later parts of this question, so if you do not know how to calculate it, make a common sense guess*.)

3(b). (4 points) The next year, 100 new students are randomly assigned to one of two teaching methods, with 50 students in each group. What are the null and alternative hypotheses for this study?

3(c). (5 points) Assuming that the standard deviation of class scores is the same from year to year, how large does the difference between the two groups' average scores need to be in order for you to reject the null hypothesis with $p < 0.05$? (*We recommend using the figure below, which was taught in class, and which shows the possible values of the mean score of those taught the new way minus the mean score of those taught the old way, assuming that there is no real difference between the groups. The "s" is equal to the standard error of the mean that you calculated in part (a).*)



3(d). (5 points) Alter or redraw the figure in part (c) above to show what would happen if there is an average 2 percentage point difference between the groups (with the new method doing better). Describe what changes in the figure.

3(e). (5 points) Using your answers from parts (a) and (d), what is the power of this study to detect a 2 percentage point difference between the groups?

4. Your father is a heavy smoker, and despite repeated attempts, he has never been able to quit. He shows you an article talking about how heavy smokers should get regular CT scans to screen them for lung cancer. Your father knows you have taken an Evidence-Based Medicine class and so he asks you for advice about whether he should believe the article and get screened. What do you need to know before you can advise him whether to get screened or not? In particular:

4(a). (6 points) The article says that those screened were 50% likely to be still alive 5 years after diagnosis, while those who weren't screened were only 15% likely to be still alive 5 years after diagnosis. Do you find this data persuasive? Why or why not?

You search the internet and find details on the main randomized control trial done on lung cancer screening. 10,000 heavy smokers were screened and another 10,000 heavy smokers were in the control group and not screened. Five years after screening, 250 had died of lung cancer in the unscreened group and 200 died of lung cancer in the screened group. State this benefits in terms of

4(b). (3 points) relative risk reduction

4(c). (3 points) absolute risk reduction

4(d). (3 points) number needed to screen to save one life

4(e). (6 points) Name three possible *harms* of screening that you would like to know more about before your father makes a decision.

5(a). (6 points) You are doing research on a hypothesis that you think has about a 20% chance of being correct. You find evidence for your hypothesis with $p < 0.05$. You estimate your power to be 50%. Now what is the probability that your hypothesis is correct?

5(b). (5 points) One of your closest competitors did exactly the same experiment, with the same power. She also got $p < 0.05$, and her work is published at exactly the same time as yours. Does this increase the probability that your hypothesis is correct relative to your answer to part (a)? Why or why not?

5(c). (5 points) What else would you like to know that might change your confidence in your hypothesis?

6. (6 points) Explain the difference between the following two placebo effects: spontaneous improvement and regression to the mean.

7. (8 points) The table below shows anonymous details for patients enrolled in a new placebo-controlled clinical trial for a drug to treat type II diabetes. The severity of their diabetes has been measured at the beginning of the trial, and will be measured again at the end in order to determine whether the drug is effective. Describe how you would randomize these patients into two parallel groups according to a matched control scheme. Explain how you would do the matching, and the reason(s) for your choice.

Patient	**Sex**	**Age**	**Race / ethnicity**	**Disease severity index at the beginning of the trial**
1	M	72	White / Caucasian	96
2	M	51	African American	97
3	F	56	Hispanic	85
4	F	75	African American	20
5	M	52	Native American	75
6	F	65	White / Caucasian	75
7	M	63	White / Caucasian	82
8	F	59	White / Caucasian	22

Ecology 379: Evidence Based Medicine **Name:**
Spring 2015
Final Exam

1a. (6 points) Draw lines between the two columns to match up terms that describe the same things. Note that "extreme" data means data showing a pattern highly suggestive of the alternative rather than the null hypothesis.

Probability(drug works|data that extreme)	posterior probability / positive predictive value
Probability(data that extreme|drug doesn't work)	likelihood
Probability(data that extreme|drug works)	*p*-value
Probability(exactly that data|hypothesis)	power

1b. (2 points) Which one should be of MOST interest to patients, practicing doctors, and members of the general public?

1c. (2 points) Which one do you need to use Bayes' Theorem to calculate, as an output?

1d. (2 points) Which one or more do you need as inputs to your Bayes' Theorem calculation?

1e. (2 points) If the drug works, which of the following *p*-values could happen? Circle all that apply.

0.001
0.04
0.06
0.1
0.2
0.5
0.99
1.5

2. You are conducting trials on a drug to treat depression. You currently enroll patients only when they meet strict eligibility criteria. You are concerned that your next study won't have enough statistical power. You have two options to increase your power, which each cost the same amount of money.

i- You could relax your eligibility criteria, ***doubling the number of patients that you study*** without needing to put any more resources into recruiting.
ii- You could tighten your eligibility criteria even further, ***reducing by half the standard deviation among patients***. Instead of spending money studying more patients within the trial, you spend the money advertising for and interviewing more potential patients, and then rejecting most of them as not fitting the eligibility criteria.

a. (6 points) Which of these strategies is most effective in increasing your power (or are they the same) and why?

b. (6 points) Which of the strategies above is better suited for studying efficacy? Which strategy is better suited for effectiveness? Explain why.

3. Consider this message from a 2011 promotional campaign run by the breast-cancer nonprofit Susan G. Komen for the Cure: “Early detection saves lives. The 5-year survival rate for breast cancer when caught early is 98%. When it’s not? 23%.” Assume that these 5-year survival rates are correct.

a. (6 points) What is “lead time bias” and why does it destroy the logic that “early detection saves lives”? Feel free to use a well labeled diagram to illustrate your answer, in addition to a description in words.

b. (5 points) What is “overdiagnosis” and how does it undermine the logic used to back up the claim that “early detection saves lives”?

c. (4 points) What data, from what kind of study, would you like to see instead of these two 5-year survival rates, before you decide whether a screening program for early detection actually saves lives?

4. (10 points) Consider the true story of Sally Clark. The frequency of cot death (where an infant dies in its crib unexpectedly and for no apparent reason) in an affluent non-smoking family like Sally's is 1 in 8500. Sally claimed that both of her two sons died in infancy of cot death. In court it was argued that this was unlikely to be true, indeed that the probability that Sally's story about cot death is true could be calculated as $1/8500 \times 1/8500 = 1/73$ million. On this basis, a jury convicted Sally of murdering her children. Describe TWO mistakes in the court's reasoning about probability.

5. A man is stopped at random on the street because he is wearing a distinctive hat beloved of the neighborhood gang of drug dealers. Ninety-eight per cent of the gang wear the hat but only 5 per cent of the local population do. Only one in 1,000 out in the street is in the gang.

a. (2 points) When deciding whether he is a member of the gang, what is the "null" hypothesis, and what is the alternative hypothesis?

b. (4 points) State the rate of **false positives** in the form "Prob(A|B)=x", replacing "A", "B", and "x" with appropriate statements or numbers.

c. (6 points) What is the probability that the man in the hat is a member of the gang?

6a. (5 points) Why do randomized trials give a placebo (e.g. a sugar pill) to patients in the control group? In other words, in what way is a trial done with a placebo control better than a study done with a no-treatment control?

b. (5 points) What kinds of placebo effects appear just as much in a no-treatment group than in a placebo group?

c. (5 points) What information could you learn from a study that had BOTH control groups, one receiving a placebo, and one receiving no treatment at all?

7. Consider the drug facts box below, which uses real numbers for a real drug. Abilify was originally FDA-approved as an antipsychotic to treat schizophrenia, but is commonly taken for depression, and is now FDA-approved as an adjunctive treatment for depression. After studying the drug facts box, answer the questions on the next page.

M020210

ABILIFY
(aripiprazole)
for inadequately treated depression in adults

OM 4

Drug Facts

What difference did ABILIFY make?	Antidepressant + **PLACEBO** (no drug) 190 people	Antidepressant + **ABILIFY** (10 mg each day) 191 people
How did ABILIFY help?		
People had lower depression scores – their score improved by 3 points more than placebo (on a scale from 0 to 60)	**6 points better**	**9 points better**
More people "no longer depressed" based on score 10% no longer depressed due to drug)	**15%**	**25%**
What were ABILIFY's side effects?		
Serious side effects:		
More people developed severe restlessness that makes it hard to keep still (21% had restlessness due to drug)	**4%**	**25%**
Symptom side effects:		
More had insomnia (6% had insomnia due to drug)	**2%**	**8%**
More had fatigue (4% had fatigue due to drug)	**4%**	**8%**
More had substantial weight gain (4% had weight gain due to drug)	**1%**	**5%**

Warning

Young adults taking anti-depressants for major depression have a risk of suicidal thinking and behavior. This category of drugs causes uncommon but serious side effects: neuroleptic malignant syndrome (fever and high blood pressure which can lead to death), tardive dyskinesia (uncontrollable facial and body movements), diabetes, low white blood cell levels and seizures.

←

7. Summarize the information on the Abilify fact box on the last page, i.e. repeat the most salient facts in your own words, for each of the scenarios below.

7a. (6 points) You are a sales rep trying to sell Abilify. Do not tell any lies, but you want to spin the facts in a favorable light.

7b. (6 points) You are a doctor trying to figure out for yourself whether Abilify is something you want to recommend to your patients.

8a. (6 points) What does the standard deviation measure, and what does the standard error measure? What is the difference between them?

8b. (4 points) If you draw a graph to summarize the results of a study, which one do you use for the error bars, and why?

Evidence-Based Medicine Schedule of Classes and Due Dates, Spring 2014

Date	Topic	Taking the Medicine and other reading to be completed BEFORE listed class date	Homework due. In addition, short comments/questions on assigned reading are to be completed on small cards at the beginning of each lecture.	Journaling or other online activity due before class
Jan 16	Introduction to course and to clinical trials	For fun, read www.bmj.com/content/319/7225/1618		
Jan 21	Childbed fever and a "natural experiment"	Prologue, Ch 1		QRQ pre-test
Jan 23	Checklists and handwashing	www.newyorker.com/magazine/2007/12/10/the-checklist		probability reflection Q1
Jan 28	Introductory probability	Ch 2		Online confirmation bias module
Jan 30	Probability continued, binomial distribution	Ch 3		
Feb 4	Hypothesis tests and the lady tasting tea	Ch 4, 5	Homework 1	probability reflection Q2
Feb 6	lading tasting tea part II	Lady Tasting Tea reading. Bring the reading to class.		
Feb 11	Likelihood ratio test	Ch. 6, 7		
Feb 13	Finish likelihood ratio test in preparation for homework 2. Start on Type I and Type II errors	Ch. 8		
Feb 18	Power analysis. Type I type II errors continued with role playing for homework 3	Ch 9, 10		
Feb 20	Reanalyze Colebrook's data. Begin continuous data	Ch 11-13	Homework 2	
Feb 25	Continuous data, mean and variance. Start on math of power	research ethics p. 1-8	Homework 3	probability reflection Q3
Feb 27	Math of power	research ethics p. 9-17		
Mar 4	Randomization	Ch 14		

Mar 6	Study design in preparation for homework 4. Breakout groups to discuss your ideas for proposals	Ch. 15		probability reflection Q4
Mar 11	Reductionism	Ch. 16		
Mar 13	Reductionism continued. Start on placebo effects, regression to the mean	www.nature.com/nrd/journal/v2/n2/abs/nrd1012.html	Homework 4 ALSO Describe a hypothesis suitable for this class to test in a trial. Give appropriate detail about randomization, controls, matching, endpoints etc. Bring FOUR hard copies of this proposal to class and ALSO submit the proposal electronically via dropbox.	
	SPRING BREAK		By the end of class on Mar 13, you will be given 3 proposals written by other students. Read and comment on these 3 proposals for ethics, feasibility and interest. Spend ~20 minutes on each. Also give them a rank of best (1) to worst (3). Your written comments and ranks are due via dropbox on TUESDAY Mar 18.	
Mar 25	Discuss the short-list and take a vote on which ones the class will do. Form groups to work on the winning proposals	Ch 17	We will distribute a short-list of highly-ranked proposals, which you need to read before class.	
Mar 27	Organizational planning in groups to carry out the trial in time to write it up before the end of semester. Placebo effects.	Ch 18		
Apr 1	Finish placebos for homework 5.	Ch 19		
Apr 3	Health systems, health insurance, intro to essay topic	Ch 20		
Apr 8	FDA and the drug approval process	Ch 21	homework 5	
Apr 10	Drug companies	dx.doi.org/10.1371/journal.pmed.0040150 Optional suggested book: The Truth About the Drug Companies by Marcia Angell		
Apr 15	Where innovative drugs come from	www.nature.com/nrd/journal/v11/n3/full/nrd3681.html	DRAFT Essay: "How do you want the US healthcare system to work?" Guideline: 750-1500 words. Longer will not necessarily	

			be graded as better: be concise, with no unnecessary words. Double spaced please.	
Apr 17	Conditional probability for probability reflection Q5. Bayes' theorem for homework 6.	Ch 22-23		
Apr 22	Screening programs, base rates, and false positives. Lead time bias and length bias.	Ch 24	You will get your drafts back	probability reflection Q5
Apr 24	Overdiagnosis. Harms vs. benefits of mammography.	Schwartz and Woloshin drug facts box	homework 6	
Apr 29	Relative and absolute risks. Number needed to treat. Statistical vs. clinical significance. Start on homework 7. Start on Ioannidis paper.	Ch 25. More detail in optional dx.doi.org/10.1371/journal.pmed.0020124	FINAL essay due	
May 1	Ioannidis paper continued for bonus homework 8.	http://www.nature.com/news/scientific-method-statistical-errors-1.14700. More detail in optional http://jama.jamanetwork.com/article.aspx?articleid=201218		probability reflection Q6
May 6	Evidence pyramid and how to use it. Explore evidence-based medicine search engine.		homework 7	bonus probability reflection Q7
May 13	Tuesday 3:30-5:30 pm EXAM		bonus homework 8 to be emailed at least 24h before exam	QRQ post-test

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 1

1. You are writing a research paper about whether burning fossil fuels causes global climate change, and you start out with a significant personal opinion about the subject (pick whichever side you like to answer this question). How would you go about doing background reading for your paper if

    a. you were concerned about avoiding confirmation bias

    b. you were not concerned about avoiding confirmation bias

2. Briefly outline three reasons why not all surgeons use checklists all the time.

3. If you were running a hospital, name two things you would do to increase compliance with guidelines on handwashing and checklists.

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 2

1a. Define what it means to say that the "*p*-value" in a trial is equal to 0.02. (Hint: try expressing it as a conditional probability, i.e. the probability of a particular outcome, given certain conditions, where you describe which outcome and which conditions.)

1b. To test whether the lady could tell whether milk or tea were added to the cup first, Ronald Fisher made the lady taste four cups each way, eight cups in total. She identified every cup correctly. What is the p-value for this experiment? Show your working.

2. Two scientists get into an argument. Jack is studying a new way of teaching College Algebra by randomly assigning students to classes that use either the new method or the old method. There are 1800 students, who he has assigned to 60 different 30-student sections of the class. Jill thinks he won't learn anything, because the students are so varied, in terms of high school preparation, ethnicity, study habits etc, and Jack hasn't controlled for all these variables. Jack says that this is no problem. Who is correct and why?

3. A randomized controlled trial is carried out to determine the effectiveness of a new treatment for a condition. 10 patients are assigned to the treatment group and 10 patients are assigned to an untreated control group. At the end of the trial, 5 untreated patients have died and 1 treated patient has died. State your hypotheses in appropriate form. Use statistics to test these hypotheses. Explain why you do or don't believe the treatment is effective.

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 3

For the following three examples, state:
a) The null hypothesis
b) The alternative hypothesis
c) A description of what a type I error looks like in this context
d) A description of what a type II error looks like in this context

1. A highway patrolman issues a blood alcohol test on a driver
2. A public school internet filter blocks suspected pornographic sites from being downloaded by students
3. A ban is proposed on a video game because of data that seems to show that it causes violence in teenagers

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 4

1. A new drug to control asthma is being tested against a placebo. The study contains 1000 patients. For following study designs listed below, please explain what the study design would look like with the above experiment and describe how you would randomize the patients:

    a. Simple parallel group
    b. A crossover design

    Now suppose you first categorized your 1000 patients into the following groups:

    300 women with a household member who smokes
    200 women with no household member who smokes
    230 men with a household member who smokes
    270 men with no household member who smokes

    Explain what the following study designs would look like (do as above):

    c. A design that controls for smoking in the household
    d. A design that controls for gender
    e. A design that controls for gender AND smoking in the household

2. In clinical trials, patients have the option to stop taking their medicine at any time that they choose. Do you include data from these patients in efficacy and/or effectiveness trials and why/why not?

3. A study is done on giving men with osteoporosis human growth hormone with calcium and vitamin D3. This is thought to halt the fall in bone density and help build bone density. Bone density was measured by a continuous Z-score. The study contains a total of 100 patients and is designed so that alpha=0.05, and power=0.8 for an effect size in which Z changes by 1.5 units between the treatment and the placebo groups. How many patients would they need in the study to make power=0.8 for change of 0.5 units in Z?

4. The drug modafinil is approved to treat narcolepsy, and has become popular off-label as a “cognitive enhancer”. The evidence so far on the efficacy and effectiveness of modafinil is mixed, and results may depend on which population is taking it and what sort of cognitive tasks they want enhanced. There is also insufficient data on side effects, especially over the long term, although most users report only minor issues, if any. You are organizing a trial to find out whether modafinil helps college students get better grades.

4a. Describe how you would design this trial using a parallel groups. Specifically, imagine you have 100 students enrolled in the trial, numbered 1 through 100 in your double-blind protocol. How many do you randomize to take modafinil, when, and for how long? What outcome do you measure?

4b. Describe how you would design this trial using a crossover design.

4c. Which design is better for measuring long-term side effects, and why?

4d. Which design has higher power, and why?

4e. How would “matching” increase the power? Which factors would you choose to match for? Pick one factor, and explain how you would perform matching within the parallel groups design.

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 5

1a. What is the difference between an active and an inactive placebo?

1b. What is the advantage of using an active rather than inactive placebo as part of a randomized trial? Give a real or hypothetical example.

1c. Give a real example of using an active placebo *in regular clinical practice outside of any randomized trial*. What is the disadvantage of giving an active rather than inactive placebo in this context?

2. Describe 4 substantially different reasons why a randomized trial might see clinical improvements in a control group taking a placebo.

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 6

1. One in every three American women has chronic hypertension. An automated blood pressure machine gives a rough indication of blood pressure at that moment in time. 83% of people with hypertension and 24% of people without hypertension are shown to have high blood pressure by the automated blood pressure machine. A woman checks her blood pressure at this machine while waiting for her prescriptions to be filled. The machine says that she has high blood pressure. What is the probability that she actually has chronic hypertension?

2. A woman is worried that she might be pregnant (she doesn't want to be), and so she uses a home pregnancy test. Ten percent of women who are worried enough about being pregnant to take a test turn out, in the end, to be pregnant. Pregnant women are 85% likely to test positive. Women who are not pregnant (including those who were very briefly pregnant but have already had an early miscarriage and so are no longer pregnant at the time of taking the test) are 2% likely to test positive.

2a. If the woman tests positive, how likely is it that she is pregnant?
2b. If the woman tests negative, how likely is it that she is pregnant?

3. Approximately 19.3% of the U.S. adult male population smoke cigarettes. If an adult male smokes, he has a 17.2% chance of getting lung cancer. If an adult male does not smoke, he has a 1.3% chance of getting lung cancer. If a man has lung cancer, what is the probability that he smokes cigarettes?

ECOL 379: Evidence Based Medicine
Spring 2014
Homework Problem Set 7

1. A Phase III clinical trial shows that a new drug reduces the frequency of strokes in high risk patients. In the trial, 5% of the control group and 2% of the treatment group had a stroke within five years. What are the relative risk reduction and the absolute risk reduction? What is the number needed to treat in order to prevent one stroke?

2. A once-untreatable type of cancer only has a 40% survival rate, but luckily a new treatment is developed. In a randomized, placebo-controlled clinical trial, treatment A increased survival from 40% with the placebo up to 80% with treatment A, with $p<0.05$. Express the effect of treatment A on survival, in the form of

2a. relative risk of death

2b. absolute risk of death

2c. number needed to treat to prevent one death

2d. Treatment B has been found to increase the survival rate from 40% with the placebo up to 90% with treatment B, with $p<0.05$. Express, in the form of absolute risk, the difference in survival rates from using treatment B rather than treatment A.

2e. The data on treatments and A and B came from two entirely separate randomized, placebo-controlled clinical trials, conducted by two different research groups. You decide to do a new randomized trial with two groups, one group receiving treatment A, the other treatment B. The previous trials on treatments A and B each enrolled 100 patients. These previous studies had an appropriate amount of power. How many patients would you plan to enroll in your new comparative effectiveness trial?

2f. Your new trial is completed as planned, and then the double-blinding is lifted. In your new trial, the survival of patients taking treatment B is higher than for treatment A ($p<0.05$), but in absolute terms it is only 85%, down from 90% in the earlier trial. Treatment A had an 80% survival rate in both your new trial, and in the old placebo-controlled trial. How might "regression to the mean" explain the apparent decline in effectiveness of treatment B, relative to the first study?

2g. Treatment B had dramatically more severe side effects than treatment A, and patients hated it. Based on results from the second trial, as a doctor, how would you decide which treatment, A or B, to give your next patient?

3. A meta-analysis of antidepressant drugs found that patients taking the drug scored about 2 points higher on a 62 point depression scale than patients taking a placebo, with $p<0.05$. Which part of the preceding sentence refers to statistical significance, and which part refers to clinical significance? Are the usual thresholds for significance met in this example?

ECOL 379: Evidence Based Medicine
Spring 2014
Homework #8

1. A scientist tests 100 drugs on a cancer cell line to test whether they stop cell proliferation. 20 of them really work, while 80 are ineffective. The scientist sets a type I error rate of 0.05. This leads to a power of 0.8 for the drugs that really work. Out of the 100 drugs tested, how many type I and type II errors are expected in the scientist's results?

2. Consider a meta-analysis of many studies of the same treatment. Although each of the studies enrolled too few patients to be conclusive, if you put all the results together you have a power of 80% to detect the drug's effects. You suspect that the prior probability that the drug is effective is around 30%. The meta-analysis shows that the drug is effective, with $p < 0.05$.

a. If there is no publication bias or other biases, what is the probability that the drug really is effective?

b. If publication or other biases mean that there is a 20% chance of inappropriately deeming a drug to be effective when in the absence of bias it would have tested negative, what is the probability that the drug really is effective?

3. Across the approximately 6000 hospitals in the United States, patients infected with flesh-eating bacteria are 25% likely to die. But in one hospital, after 8 patients were recently infected, they all died. What is the probability of this happening? Does this seem suspicious to you? If you were administering an insurance program (whether private or government) that imposed financial penalties on hospitals for poor performance, would you fine this hospital?

Evidence-Based Medicine

- Instructor: Joanna Masel (office hr Wed 11:30-12:30)
- TA: Parris Humphrey (office hr Mon 2:30-3:30)
- Lecture schedule, reading assignments, office hours etc. are all on D2L: please read.
- If you haven't, order the book "Taking the Medicine".
- Make sure you forward your D2L email so you get any class announcements

Not your average class

- Compulsory reading
  - How many of you regularly read books?
- Lots of discussion in class: ppts will be posted on D2L, but may be minimal, or we may stray from them, you should take your own notes on what you think is important
- Critical thinking: understanding not remembering
- Problem solving: when you don't know the answer, try drawing a picture, or answering a simpler question that seems related, or making up a more concrete and maybe simpler example that is easier to think about

Readings

- There will be a quick quiz on the reading at the start of every class STARTING NEXT CLASS. Bring index cards to class with you, and write the answer and your name during the first 5 minutes of class.
- No need to fill the space, concise answers welcome
- If you will be missing class for a valid reason (eg med school interview), please let me know a week in advance so you can take the quiz early.
- Quizzes cannot be made up late. The lowest 2 are dropped. If you are sick for more than 1 class AND have a doctor's note, please talk to me and we will arrange something.
- Next reading is from the book, the following reading is an article: always check D2L schedule for the latest information

Attendance is compulsory

- Not just about doing the quiz
- For many classes, the ONLY way to catch them up effectively is in office hours
- Talk to me in advance for med school interviews etc., and after the fact in case of illness.
- Avoidable absences will not be excused (although we will still try to help you catch up during office hours)

Grading

- Cards handed in each lecture (2 may be dropped) 10%
- Participation in class discussions 5%
- 8 Homework problem sets 16%
- 1 online homework module 1%
- Participation in required journaling 3%
- Research proposal 12%
- Written comments on the proposals of others 3%
- Leadership and participation in class experiment, its analysis and writeup 5%
- Essay 15%
- Final exam (short answer and statistical questions) 30%

Class emphasizes "ah-ha" moments and transfer to real-world scenarios

- For this reason, there is only one, final exam
- If you don't understand something in class, do NOT wait until the final exam to try to catch up: come to office hours ASAP and we will help you.
- Many questions on the problem sets come from past exams

Assessments

- You will need to take two assessments for this class:
  - Attitudes Toward Statistics: passed out in class today
  - Quantitative Reasoning Quotient: This is available on D2L under the quizzes section
  - Please finish both by class next Tuesday (1/21)

Journaling (3% of grade)

- AFTER you take the D2L quiz, go to D2L/Discussions and you will see a thread called "probability reflections"
- Add your first journal entry to answer the following question:

What does it mean to say that
"the probability of dying from disease X is 20%"?

Classroom behavior

- 5% of your grade comes from participation in class discussions. You must be present on-time, awake and involved.
- Narcoleptics (with a medical note) will receive a disability accommodation excusing them from nodding off, but NOT exempting them from being awake often enough to actively participate in discussions.
- Computers, phones etc. allowed in class ONLY if you are looking up something relevant to the point under discussion, with a view to share it with the class.
  - See me before the class in question if you need an exemption on the grounds of a personal standby emergency

What do doctors do?

1. Diagnose
2. Treat (for better or worse)
   - Save lives (or kill)
   - Improve lives (or make them worse)
   - Include preventative measures
3. Reassure, provide ritual

Are all efforts to save or improve lives medicine?

Better food and shelter saves lives.

Music improves lives

Is unhappiness a disease?

Assume the diagnosis is correct. How does a doctor know which treatment will help?

1. The last person he tried it on got better
   - Maybe they would have got better anyway
2. That's what the doctor learned in med school
   - How do the med school profs know?
3. The treatment makes sense. Eg, if you have a tumor, you should cut it out.

Is "Western" medicine is better than "alternative" medicine? Why or why not?

- Western medicine =
  - things taught at medical schools and administered by MDs / nurses / physical therapists
- Alternative medicine =
  - anything else
  - eg, homeopathy, naturopathy, acupuncture, chiropractice, meditation, prayer

Clinical trial

- Take lots of people with the same disease. Give only half of them the treatment, and see if they do better.
- Seem simple?

Randomization

- Not all patients are the same, even if they have the same diagnosis
- Does this matter? Do we need to understand all the differences?
- More understanding → more diagnoses
- If we randomly put the patients with the same diagnosis into treatment groups, and one group does better, then a new RANDOM PATIENT WITH THAT DIAGNOSIS is best off getting that treatment
- Our diagnoses aren't perfect, but the rest of the variation is present equally in both groups, so it averages out

If the effect is large and fast, you may not need a randomized trial

- If a drug takes away all your pain, or makes you vomit, or induces an abortion, or does something else dramatic, you can figure it out
- The dramatic effect (eg vomiting) may not help your disease
- If the dramatic effect is delayed, it may be hard to pin down exactly what caused it

Alternatives to the evidence-based medicine of clinical trials

Basis for clinical decisions	Marker	Measuring device	Unit of measurement
Evidence	Randomized controlled trial	Meta-analysis	Odds ratio
Eminence	Radiance of white hair	Luminometer	Optical density
Vehemence	Level of stridency	Audiometer	Decibels
Eloquence (or elegance)	Smoothness of tongue or nap of suit	Teflometer	Adhesin score
Providence	Level of religious fervor	Sextant to measure angle of genuflection	International units of piety
Diffidence	Level of gloom	Nihilometer	Sighs
Nervousness	Litigation phobia level	Every conceivable test	Bank balance
Confidence*	Bravado	Sweat test	No sweat

*Applies only to surgeons

Clinical trials work for non-medical questions too

- Does working out every day make muscles visibly bigger?
- Are parents more likely to lend children the car if they say "please"?
- Which diets actually work?
- Science doesn't have to be about arcane topics that are hard to grasp. Everyday life can be subjected to the same method.
- Anything with
  1. a starting population that can be randomized into groups
  2. an intervention
  3. an outcome measurement

If science is that easy, why don't we understand more?

- Psychological resistance to the method
- Need to understand statistics
- This course will spend a lot of time on statistical concepts, with only the necessary minimum on the mathematical recipes
- Capstone is to design and carry out our own trial

Class quiz

Give an example of a primitive treatment that was effective but came with serious side effects

Reminders

You should all have done the QRQ by now.
Last I looked, 7/40 people hadn't done it yet.

The probability reflection is due by next class
(under D2L/Discussions)
Online module after that
Office hours:
Dr Masel: Wed 11:30-12:30 LSS379A
Parris: Mondays 2:30-3:30

Discussion points from reading

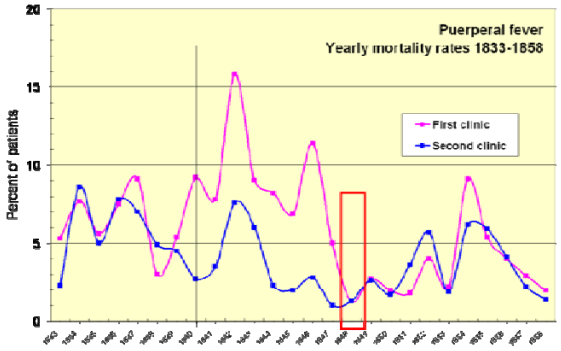
- Comments on opium?
- Comments on the inferior medicine of "primitive" cultures?
- Comments on how small a contribution medicine has made to improved health and life expectancy?

More Discussion

- Why didn't doctors notice that they did no good?
- Why didn't the patients notice?
- Have these reasons changed?

- The online module will suggest some reasons

See first three parts of Colyer C (1999). Childbed Fever: A Nineteenth-Century Mystery. National Center for Case Study Teaching in Science for next exercise



Class quiz

Describe one checklist that was discussed in the reading, and found to be effective in a medical setting. When should it be used?

To insert a catheter

1. Wash your hands
2. Clean patient's skin with chlorhexidine
3. Put sterile drapes over the entire patient
4. Wear sterile mask, hat, gown, gloves
5. Put a sterile dressing over the catheter site once the line is in

- All 5 were supposedly already standard, but nurses counted that for more than 1/3 patients, at least 1 step was skipped
- Infection rate after having a catheter line for 10 days went from 11% to ZERO the next year, and only 2 infections during the next 15 months
- In that one hospital in that short time, the checklist prevented 43 infections and 8 deaths, and saved $2,000,000

Patients on ventilators

- Head of their bed should be propped up by at least 30 degrees so oral secretions don't get in the windpipe
- Antacids
- Occurrence of pneumonia fell by a quarter
- 21 fewer patients died than the year before
- ICU doctors and nurses made their own checklists ⇒ length of stay dropped by half

Not just the piece of paper

- Hospital executives in Michigan had to visit the ward at least once a month and listen to complaints
- Chlorhexidine and drapes were not easily available at first: executives fixed this quickly
- Persuaded supplier to sell a "kit" of catheter + drapes + chlorhexidine
- Within 3 months, Michigan infection rate went down 66%

Empowering nurses

- Nurses were authorized to stop doctors if they saw them skipping a step
- Every day, nurses would ask the doctor "is that line still needed?", so they weren't used longer than necessary
- Hierarchy can kill
- Doctors are often offended when asked to use checklists

Surgeon blog quote

Certainly, there is something to be said for meticulous routines when it comes to surgery or other procedures. But do we need mandatory 19 item checklists? Why stop there? Why not make it a 40 item checklist? Why not make the attending surgeon write an essay on how to avoid complications before every case? Or how about having the surgeon and all assistants read the chapter corresponding to the proposed operation from the textbook out loud together (alternating paragraphs) prior to making the incision?

It's good to be organized and precise in surgery. Limited checklists are useful in this regard. We ought to mark our initials on the correct side of the hernia repair. Point taken. Nothing groundbreaking here. We don't want to be operating on the wrong leg or leaving sponges inside bellies. But it's rather a ridiculous leap to think that death rates can be halved just by following a series of irritating instructions on a laminated list.

Comment by surgeon on same blog

Surgical safety is always paramount when I do an operation. But to use the results of this study as definitive proof that by simply implementing Dr Gawande's 19 point checklist will save thousands of lives is misguided. This was a non-randomized, non-blinded study. It's not hard science. Long, indepth checklists are only going to complicate health care. Will we need different checklists depending on the operation? Will there be separate checklists for doctors vs nurses vs anesthesia staff? Who will be in charge of determining each checklist? A subcommittee of the AMA? A national bureaucracy? Common sense and moderation, as usual, ought to be our guiding principles.

Why do doctors hate checklists?

1. Doctors have big egos, don't like to be corrected by nurses and/or checklists
2. Too busy, too many tasks, too much bureaucracy already
3. Medical research establishment that stresses more "exciting" molecular findings, or surgeons' amazing skill, or cool prosthetic technology, not mundane research on checklists

Advantages of checklists

1. Remind you of what you know you should do, but may skip sometimes, especially in an emergency and when there are many things going on at once
2. Establish a higher standard of minimum care: not everyone knows about the evidence supporting each step

- Airline pilots use checklists. Cooks use checklists (known as recipes). Why not doctors? ICUs are complicated places
- Cultural change is key

See last part of Colyer C (1999). Childbed Fever: A Nineteenth-Century Mystery. National Center for Case Study Teaching in Science for exercise, which was continued here

Modern handwashing

- One calculation suggested that if a hospital nurse washed her hands with soap and water, according to guidelines and after every contact, she would spend 80% of her shift washing hands.
- Unsurprisingly, those guidelines on handwashing were ignored, and hospital patients died.
- Alcohol foam is faster: 10-20 s. instead of 90 s, and you can multitask
- As compliance increased, hospital infections decreased.
- Studies show MASSIVE improvements whenever compliance goes up

Handwashing today

- With aggressive promotion, training, and convenient foam dispensers at every patient door, compliance can be ~90% (100% would be better!)
- Without such measures, compliance can be low, even 10%
- Observational studies clearly show that medics wash their hands less often than they think
- Compliance may be lower in
  - doctors or nurse assistants, nurses better
  - men
  - ICU / other critical settings

Test systematically, collect data

- Imagine that 40% of patients die, but treatment reduces this to 20%. Good treatment?
- Each decide your fate by rolling a die

Data	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

Class quiz

What is laudanum? Why was it a medical advance?

Homework

Online module was due today, this completed the material you need to do homework 1. Now available online and due in class in 1 week.

Probability Reflections Question 2 is now available, and also due before class in 1 week. Answer the question as a new post, under Discussions / Probability Reflections.

Test systematically, collect data

- Imagine that 40% of patients die, but treatment reduces this to 20%. We agreed that while not perfect, this was a good treatment.
- We did a randomized controlled trial, where each of you decided your fate by rolling a die:

Data	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

Interpret the data

- We know this treatment works.
- But would the data we have be enough to persuade you?
- If everyone without treatment dies, and everyone with treatment survives, and "everyone" is a lot of people, result is obvious
- This is rare
- The rest of the time you need statistics

Three hypotheses

1. Treatment helps
2. Treatment does nothing
3. Treatment harms

- The default or "null" hypothesis is that the treatment does nothing
- Use statistics to ask if the data is enough to change your mind (towards either 1. or 3.)

Likelihood

- *If the null hypothesis is true, how likely are you to see this data?*
- *If the null hypothesis is NOT true, how likely are you to see this data?*

Data	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Probability(data|hypothesis)
- Need to understand probabilities, time for a digression

Write your answers on a piece of paper, with your name

1. What is the probability of rolling a 4?
2. If you were to roll the dice 100 times, how many times do you expect to get a 4?
3. In this class of 40 dice-rollers, how many people do you expect to get a 4?
4. Roll the dice. How many people got a 4? Do you want to change your mind to #3? Don't cross out your answer, but write a new one below and say why you changed.
5. What was the frequency of getting a 4? Is this the same as the probability?

Probabilities and frequencies

- Frequency is the number of times something happened, divided by the number of times it could have happened.

  6. What is the frequency of deaths in our treated "patients"?
- Frequencies can be used to estimate "true" probabilities, representing the limit of a very large number of things that could have happened.

  7. What is the probability of death of our treated patients?
  – "per cent" means per hundred. Probability = 0.2 = 4/20
- Probabilities must be between 0 and 1.
- If you add up frequencies or probabilities of every alternative, they add up to 1.

To calculate frequencies, you need to be comfortable with fractions and percentages

8. Without using your calculator, what is 0.1% of 10,000?

Estimate probabilities from frequencies

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

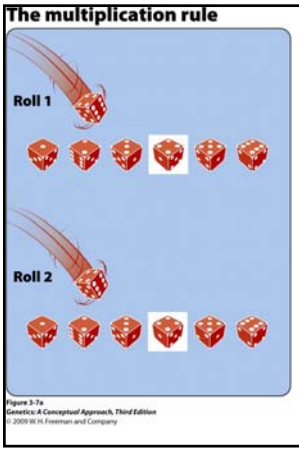
- Prob(live if treated|treatment matters) = 20/22
- Prob(live if untreated|treatment matters) =

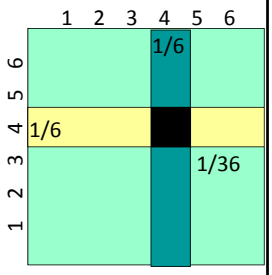
- Prob(live if treated|treatment does nothing) =
- Prob(live if untreated|treatment does nothing) =

Add to your list of answers

9. What is the probability of rolling even rather than odd?
10. We will roll twice, recording just odd vs. even.
    What is the probability of two evens?

The multiplication rule





The black square is when you have the blue AND yellow independent events

When are events independent?

- How many brothers (not sisters) do each of you have?
- Do you expect the men in this room to have fewer brothers than the women do?
- The sex of a baby is INDEPENDENT of the sex of other children in the same family.
- But if I asked "how many boys in your family (including you)?" this is not independent of whether I am asking a man or a woman. The men in the room would on average report 1 more boy in their family than the women would.

Add to your list of answers

11. What is the probability of rolling even then odd?
12. What is the probability of one even and one odd, in any order?

- EE, EO, OE and OO are "equiprobable"
- 2 evens, 1 even & 1 odd, 2 odds are NOT equiprobable, so be careful how you list the possible outcomes!

Which has higher probability (or are the probabilities the same)?

13. EEEEEE vs. EEOEOO

- EEEEEE is less "representative", but still has exactly the same probability as any other sequence.

14. The probability for each of them is

$$0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.5^6 = 0.015625$$

What are the probabilities for the different ways of rolling a multilink cube?

Up Out Down Balanced

- You can't reason based on equiprobable outcomes
- The only way to estimate the probabilities is to roll many times and use the frequencies

Results from a more patient class/teacher

Position	Number of Occurrences	Experimental Probability
OUT	4124	0.76
UP	765	0.14
DOWN	469	0.09
BALANCED	72	0.01
TOTALS	**5430**	**1.00**

Class quiz

- What made cinchona/quinine different from all previous drugs?
- Why did it fall out of favor in Rome in the mid 1600s?

What are the probabilities for the different ways of rolling a multilink cube?

Up Out Down Balanced

- You can't reason based on equiprobable outcomes
- The only way to estimate the probabilities is to roll many times and use the frequencies

Results from a more patient class/teacher

Position	Number of Occurrences	Experimental Probability
OUT	4124	0.76
UP	765	0.14
DOWN	469	0.09
BALANCED	72	0.01
TOTALS	**5430**	**1.00**

"Model" the multilink cube with an urn

- 5430 balls in the urn: 4124, 765, 469, 72 of each type
- Rolling the multilink cube is like drawing a ball at random from the urn
- Urns can be analyzed via equiprobability (classical probability)
- Multilink cube needs "frequentism" version of probability

Position	Number of Occurrences	Experimental Probability
OUT	4124	0.76
UP	765	0.14
DOWN	469	0.09
BALANCED	72	0.01
TOTALS	5430	1.00

Are patient outcomes better described by classical probability (equiprobable dice rolls or urn-drawing), or by frequentist probability (multilink cube, roll many times and see)?

Likelihood

- *If the null hypothesis is true, how likely are you to see this data?*
- *If the null hypothesis is NOT true, how likely are you to see this data?*

Data	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Probability(data|hypothesis)
- Prob(patient X lives|treatment does nothing)

$p = 32/40 = 4/5 = 0.8$

Binomial distribution: probability that *k* out of *n* survive

$$\binom{n}{k} p^k (1-p)^{n-k}$$

P(first *k* live, remaining *n*-*k* die) =

The *k* survivals out of *n* could come in any order, this is just the probability that it is the first *k* that survive. How many ways to get the same result?

How many ways are there to pick *k* patients out of *n*?

"*n* choose *k*" $\binom{n}{k}$

Examples of *n*-choose-*k*

How many ways are there to pick 1 patient out of 2?

2

How many ways are there to pick 2 patients out of 4?

Patients ABCD, you could pick

AB, AC, AD, BC, BD, CD = 6 ways in total

How many ways are there to pick 2 patients out of 5?

AB, AC, AD, AE, BC, BD, BE, CD, CE, DE = 10

Which do you think is bigger: 10-choose-2 or 10-choose-8?

Now calculate them from patients ABCDEFGHIJ

Always in pairs:

choosing AB is the same as choosing CDEFGHIJ

10 ways of picking the first patient

9 ways of picking the second

Each pair can come in either order, so divide by 2

10×9/2 = 45

"Availability" heuristic

It is easier to think of sets of 2 patients than sets of 8 patients

The availability heuristic is a trap in medical diagnoses, be careful of always choosing the diagnosis that comes to mind most easily

How many ways to choose 4 patients out of 8?

8 ways to pick the first patient

7 ways to pick the second

6 ways to pick the third

5 ways to pick the fourth

But now we have counted the same patient sets multiple times, ABCD, BACD, ABDC, DCAB etc.

How many times have we counted the same 4 patients?

4 choices for which patient is counted 1st

3 for which is counted 2nd

2 for 3rd, 1 for 4th

Answer: 8×7×6×5/(4×3×2 ×1) = 70

Slightly different question

Instead of a patient being "living" or "dead", imagine that each patient has a unique outcome. With *n* patients, *n* unique outcomes.

In how many different orders might you see these *n* unique outcomes?

n choices for what comes first

n-1 for what comes second

n-2 for what comes third

…

n! = *n*(*n*-1)(*n*-2)(*n*-3)…3 × 2 × 1

Binomial distribution: probability that *k* out of *n* survive

$$\binom{n}{k} p^k (1-p)^{n-k}$$

There are *k* survival outcomes and *n*-*k* death outcomes. How many different orders are there for these outcomes?

If we consider each patient outcome as unique, they are *n!* ways to order them

By how much does this overcount the possible orders of patient outcomes when we sum up what matters as simply "living" vs. "dead"?

We counted every possible order for the *k* surviving patients, so we overcounted by *k!*

Similarly, we also overcounted the deaths by (*n*-*k*)!

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

8-choose-4

$$\binom{8}{4} = \frac{8x7x6x5}{4x3x2x1} = \frac{8!/4!}{4!} = \frac{n!}{k!(n-k)!}$$

Class quiz

How could willow bark (which did not cure malaria) gain favor over Cinchona bark (which does cure malaria) as a treatment for malaria?

Discussion points from reading

- Different meanings of the word "experiment"
  - To try out
  - To test systematically
- How often does a treatment need to work before you believe it?

Lady Tasting Tea

- Break up into groups of THREE students
- Act out the dialogue given in the handout
- Written for 2 men and 1 woman, but adjust as needed, eg
  - model and escort could both be gay men
  - older gentleman could be an older lady
- Once you are done, work on the questions in your group. Make notes, you will need present your answers to the class, AND you will work in the same group on Thursday and need your notes then.

Homework

- Read Fisher's paper, download from the D2L schedule
- It was written in the 1930s, you might it hard going: you CAN do it, but allow extra time and give yourself permission to read slowly
- Prior knowledge of Fisher's paper is essential for Thursday's class

Class Quiz

How many cups of tea does Fisher recommend that the tea lady should taste? What is the probability of guessing that number of cups correctly by chance alone, even if the lady cannot really tell the difference?

Answer questions in part II

Work in the same groups of 3 as last lecture, and then we will compare notes on each question

Let me know when your group gets up to Q8, then skip to Q9

Definitions of experimental design (for Q7)

"A study design used to test cause-and-effect relationships between variables. The classic experimental design specifies an experimental group and a control group. The independent variable is administered to the experimental group and not to the control group, and both groups are measured on the same dependent variable. Subsequent experimental designs have used more groups and more measurements over longer periods. True experiments must have control, randomization, and manipulation."

"A branch of statistics that attempts to outline the way in which experiments should be carried out so the data gathered will have statistical value."

"The design of an experiment refers to the structure of the experiment, with particular reference to the:

- set of treatments included in the study.
- set of experimental units included in the study.
- rules and procedures by which the treatments are assigned to the experimental units (or vice versa).
- measurements that are made on the experimental units after the treatments have been applied."

Q8: Let's do a tasting

Act out Part III, then answer questions 1-4

Skip questions 5 and 6: we will try to come back later this semester, after we have learned some more material

Class Quiz

Some patients are sicker than others. If you give one patient a drug and not the other, it is hard to tell whether it was the drug that made the difference. What did Louis suggest to avoid this problem? What did Louis actually do?

Expectations for this course

- Some of you are already asking about the final exam
- Homework problem sets are mostly drawn from past exams (although of course they are easier open book at home, straight after the topic is taught)
- What I care about (and try to write exam questions on) is whether you can take first understand the core concepts of this course, and then APPLY them to new, real-life scenarios.

Lady tasting coffee Part III Q1-4

Back to our class "experiment" on a drug

What is the probability of seeing our data

- if the treatment does work?
- if it doesn't work?

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

Estimate probabilities from frequencies

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Prob(live if treated|treatment matters) = 20/22
- Prob(live if untreated|treatment matters) = 12/18 = 2/3

- Prob(live if treated|treatment does nothing) = 32/40 = 4/5 = 0.8
- Prob(live if untreated|treatment does nothing) = 32/40 = 4/5 = 0.8

What is the probability of seeing this data in a universe where the treatment does nothing?

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Prob(patient X lives|treatment) = 0.8
- Prob(patients X and Y live) = 0.8^2
- Prob(32 patients in a row live) = 0.8^{32}
- Prob(first 32 patients live, next 8 die) = $0.8^{32}0.2^8$

Binomial distribution:
probability that *k* out of *n* survive $\binom{n}{k} p^k (1-p)^{n-k}$

There are *k* survivals and *n*-*k* deaths. How many ways to choose *k* names for the "life list"?

There are *n!* different orders for writing down all *n* names.

The first *k* on each list survive. For each set of *k* survivors, how many of the *n*! lists include those same *k* survivors?

The *k* survivors are written down in *k*! different orders at the top

Similarly, the death names are written down in (*n*-*k*)! ways at the bottom

Total number of lists divided by how many times you have the same list in disguise = $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

What is the probability of seeing our data
in a universe where the treatment does nothing?

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Prob(20 live and 2 die) = $0.8^{20}\times0.2^{2}\times22!/(20!2!)=0.107$
- Prob(12 live and 6 die) = $0.8^{12}\times0.2^{6}\times18!/(12!6!)=0.004$
- Prob(exactly this data) = $0.107\text{x}0.004=0.000433$

Class Quiz

How did the invention of the gaslight help the synthetic dye industry get started?

Discussion Points From Reading

- Selectiveness of dyes made it conceivable that a chemical could have the selective power to kill

Binomial distribution: probability that *k* out of *n* survive

$$\binom{n}{k} p^k (1-p)^{n-k}$$

There are *k* survivals and *n-k* deaths. How many ways to choose *k* names for the "life list"?

There are *n!* different orders for writing down all *n* names.

The first *k* on each list survive. For each set of *k* survivors, how many of the *n*! lists include those same *k* survivors?

The *k* survivors are written down in *k*! different orders at the top

Similarly, the death names are written down in (*n-k*)! ways at the bottom

Total number of lists divided by how many times you have the same list in disguise = $\binom{n}{k} = \frac{n!}{k!(n-k)!}$

What is the probability of seeing our data in a universe where the treatment does nothing?

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Prob(20 live and 2 die) = $0.8^{20}\times0.2^{2}\times22!/(20!2!)=0.107$
- Prob(12 live and 6 die) = $0.8^{12}\times0.2^{6}\times18!/(12!6!)=0.081$
- Prob(exactly this data) = 0.107x0.081=0.0087

In contrast, what is the probability of seeing this data in a universe where the treatment works?

	Treat	Don't treat	Total
Live	20	12	32
Die	2	6	8
Total	22	18	40

- Prob(20 live and 2 die) = $(20/22)^{20}\times(2/22)^{2}\times22!/(20!2!)=0.284$
- Prob(12 live and 6 die) = $(2/3)^{12}\times(1/3)^{6}\times18!/(12!6!)=0.196$
- Prob(exactly this data) = 0.284 x 0.196 =0.0557 > 0.0087
- Is this difference big enough to persuade us?

Is this difference big enough to persuade us?

- Prob(data $|p$ =0.8) = 0.0087
- Prob(data $|p_t$=20/22, p_u=2/3)= 0.0557
- If you account for male/female, young/old, fit/unfit etc., eventually you can explain EVERY PATIENT and have p_{data}=1
- Each factor you account for, even if it is NOT important, will artificially increase p_{data}
- We are comparing accounting for zero factors to accounting for one factor (treatment)
- Using 2 estimated parameters (20/22 and 12/18) instead of 1 (32/40) takes us slightly closer to our absurd p_{data}=1 perfect fit
- There is one degree of freedom (df) difference
- Is the difference more than we would expect given 1 df?

Likelihood ratio test (G-test)

- Likelihood(data|p= 0.8) = 0.0087
- Likelihood(data|p_t=20/22, p_u=2/3) = 0.0557
- Likelihood ratio L = 0.00277/0.000433 = 6.4
- G = 2 ln L = 3.7
- If the null hypothesis is true, G has a χ^2 distribution with the appropriate number of dfs.

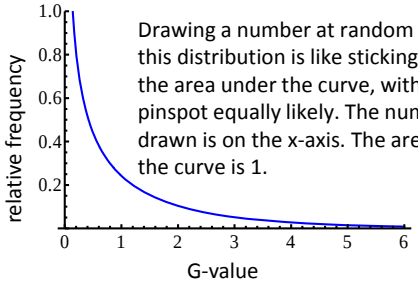
Table 3.4 Critical values of the χ^2 distribution

	P								
df	0.995	0.975	0.9	0.5	0.1	0.05	0.025	0.01	0.005
1	0.000	0.000	0.016	0.455	2.706	3.841	5.024	6.635	7.879
2	0.010	0.051	0.211	1.386	4.605	5.991	7.378	9.210	10.597
3	0.072	0.216	0.584	2.366	6.251	7.815	9.348	11.345	12.838
4	0.207	0.484	1.064	3.357	7.779	9.488	11.143	13.277	14.860

3.7

Chi-squared distribution (1 df)

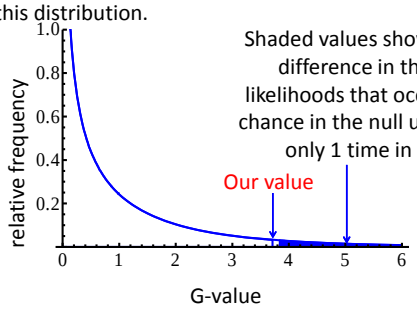
For a drug that DOESN'T work (null hypothesis), each experiment like the one we did in class generates a value from this distribution.



Drawing a number at random from this distribution is like sticking a pin on the area under the curve, with each pinspot equally likely. The number drawn is on the x-axis. The area under the curve is 1.

Chi-squared distribution (1 df)

For a drug that DOESN'T work (null hypothesis), each experiment like the one we did in class generates a value from this distribution.



Shaded values show a big difference in the 2 likelihoods that occurs by chance in the null universe only 1 time in 20

Is this data enough to persuade us to reject the null hypothesis?

- If the null hypothesis is true, the probability that difference in likelihood would be that big is a little greater than 0.05 (0.054, to be precise)
- Cannot reject the null hypothesis
- No evidence that the treatment works, even though the true effect is dramatic enough (death rate down from 40% to 20%) that we would desperately want a treatment like this if we were sick.

Homework problem set 2 is available on D2L and due in 1 week

Two different kinds of error

		Null hypothesis	
		True (no diff)	False (difference)
Decision	Accept null (no diff) p > 0.05	Correct	
	Reject null (diff) p < 0.05		Correct

Type I vs. Type II errors

- Type I errors are false positives,
  type II errors are false negatives
- Type III error: not being able to remember which is type I and which type II
- When are you most concerned with type I errors (false positives)? With type II (false negatives)?

Power

- If the treatment does something, how likely are you to notice?
- Power = 1 – type II error = 1-β
- α = type I error (*p*-value cutoff, eg 0.05)
- High power if

Class Quiz

Why did it take so long to realize that Antifebrin caused liver and kidney damage?

Discussion Points from Reading

- Paracetemol = acetaminophen = Tylenol

Two different kinds of error

		Null hypothesis	
		True (no diff)	False (difference)
Decision	Accept null (no diff) p > 0.05	Correct	Type II Error β False negative
	Reject null (diff) p < 0.05	Type I Error α False positive	Correct (Power=1-β)

Power

- If the treatment does something, how likely are you to notice?
- High power if
  - lots of patients
  - treatment has large effect
  - willing to put up with lots of type I errors
- Which one can/should you do something about?
- Ask a statistician "I want to be 95% sure that if my drug reduces death rates from 40% to 20%, then I will notice. How many patients will I need in my study?"
- Or learn to do it yourself!

Class Activity

- Imagine that 40% of patients die, but treatment reduces this to 20%.
- Everyone will roll their ten-sided dice 20 times, each roll representing a patient.
  - 10 treated and 10 untreated
- For the treated patients
  - If you roll a 0 or 1 the patient dies, otherwise they live
- For untreated patients
  - If you roll a 0,1,2 or 3 the patient dies, otherwise they live

Sample data

- Imagine that 40% of patients die, but treatment reduces this to 20%.
- Treat 10 patients, leave 10 untreated

Treated	3✞	2	3	2	0	3	0	1	3	5
Untreated	5✞	6	2	5	3	2	3	4	3	2

- Right direction in 6 experiments, wrong in 3, no difference in 1
- But right direction is not enough, need $p < 0.05$

Sample data, 10 patients each

Treated	3✞	2	3	2	0	3	0	1	3	5
Untreated	5✞	6	2	5	3	2	3	4	3	2
p-value	0.36	0.06	0.60	0.15	0.03	0.60	0.03	0.11	1	0.15

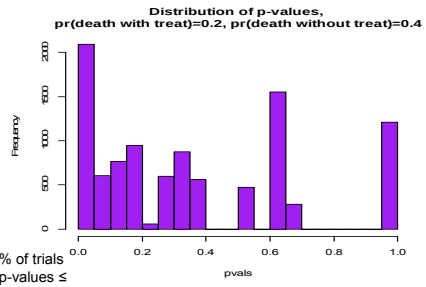
If you got different numbers from the sample data, Parris can give you a *p*-value from Excel

	Treat	Don't treat	
Die	3	5	
Live	7	5	
overall live	0.6		=(C3+B3)/20
treat live	0.7		=B3/10
no treat	0.5		=C3/10
L H0	0.04314		=BINOMDIST(B3,10,B5,FALSE)*BINOMDIST(C3,10,B5,FALSE)
L H1	0.065665		=BINOMDIST(B3,10,B6,FALSE)*BINOMDIST(C3,10,B7,FALSE)
G	0.840237		=2*LN(B9/B8)
p	0.359329		=CHIDIST(B10,1)

Warning

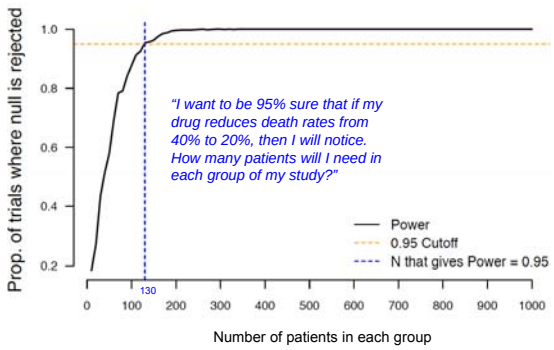
- Excel is not the best calculator in the world
- To calculate $n!/k!(n-k)!$, you calculate a very big number and divide by 2 other big numbers
- The first number, eg 100!, may be too large for your calculator
- To calculate 100-choose-10, take advantage of cancelling out
- Instead of 100!/90!, do 100×99×98×97×96×95×94×93×92×91
- Or use Wolfram Alpha instead of a normal calculator

Experiment Run 10,000 times



20.9% of trials had p-values ≤ 0.05

Power depends on the number of patients



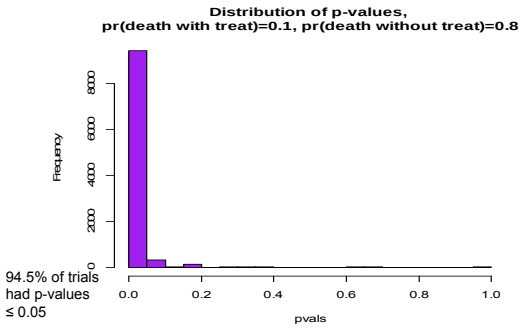
Power / type II error depends on what effect size you are looking for

- 79.1% is the type II error for a 20 patient study IF
  - 40% of patients die, but treatment reduces this to 20%
- We don't know the real effect size
- Different hypothetical effect size gives a different type II error rate for the same experiment, with the same number of patients

Class Activity II

- Imagine that 80% of patients die, but treatment reduces this to 10%.
- For 10 treated patients
  - If you roll a 0 the patient dies, otherwise they live
- For 10 untreated patients
  - If you roll a 0,1,2,3,4,5,6, or 7 the patient dies, otherwise they live
- What happens to the power compared to the previous activity?

Experiment Run 10,000 times



What if 0.8% untreated patients die, vs. 0.1% treated? Would 10 patients/group be enough?

10 patients per group enough to see (with 95% power) a reduction in deaths from 80% to 10%. What if 0.8% untreated patients die, vs. 0.1% treated? Would 10 patients/group still be enough?

- Nobody would die and you wouldn't learn anything
- Need lots of patients to spot rare events

Weird side effects are easier to spot

- Imagine taking a drug increases the rate of a very weird death from 0.00000001% to 0.01%
- Inside a trial, you wouldn't have enough patients to see
- Outside a trial, you notice this weird thing
- Now imagine a different drug increases "normal" deaths from 10% to 11% (i.e. rate in population goes from 10% to 10.01% when 1/100 people take the drug)
- Far more drug-caused deaths, but you don't notice
- Antipyrine caused rare blood disorder
- Antifebrin caused liver and kidney damage, common anyway
- Slow effects, whether rare or common, may not be seen during the timeframe of the study
- It takes a lifetime to learn about the lifetime effects of a drug

Name the type I and type II errors when testing whether a drug:

1. Works
2. Has side effects

		Null hypothesis	
		True (no diff)	False (difference)
Decision	Accept null (no diff) p > 0.05	Correct	Type II Error β False negative
	Reject null (diff) p < 0.05	Type I Error α False positive	Correct (Power=1-β)

How do you want to balance type I (false pos.) vs. type II (false neg.) errors during drug discovery?

I am willing to accept
____ useless drugs (false positives) for every
____ good drugs that we miss (false negatives)

How do you want to balance type I (false pos.) vs. type II (false neg.) errors for side effects?

- Imagine a drug that has been approved (and therefore is likely effective) as an elective treatment, but which may come with unacceptable side effects.
- When evaluating side effects, I am willing to reject
  ____ safe drugs (false positives) for every
  ____ dangerous drugs that we miss (false negatives)

Break up into groups of 4

- One person gets to be the doctor
- The others are
  2. patient for whom no other drug works
  3. drug company, interested in 12-month outlook
  4. insurance company that will pay for the drug

Now write down a new summary of how you feel in your new role, with arguments and numbers

I am willing to accept
____ useless drugs (false positives) for every
____ good drugs that we miss (false negatives)

When evaluating side effects, I am willing to reject
____ safe drugs (false positives) for every
____ dangerous drugs that we miss (false negatives)

- Persuade the doctor! He/she will make the decision.

Homework 3 on type I vs type II errors is now posted and due in 1 week

Don't forget probability reflection Q3

Class quiz

How and why did Bayer create heroin from morphine?

Discussion points from reading

- Why were the results of the serum therapy not accurate?
- John Cowan: "The days of controls are no longer possible, it is not fair to them."
- If you were one of the cheating doctors, would you have tried to give serum therapy to the healthiest patients, or to the sickest ones?

Discussion points from reading

- "Risks versus benefits"
  - Harms versus benefits
- Sulphalimide coming from the non-dye portion of the Prontosil molecule is another example of dumb luck in drug discovery, trying molecules almost at random.

Colebrook's Use of Prontosil

- Colebrook was interested in the high rates of puerperal fever in women in the maternity ward in London in the early 1930s.
- Once infected, death rates were around 20%
- After reading Domagk's report on Prontosil, he asked I.G. Farben for a supply
- First, he repeated Domagk's experiment on mice, then used it on people

Let's look at Colebrook's data on puerperal fever

- Year before: 42 deaths out of 210.
- 38 women then took the drug: 3 died
- Ignore for now the fact that these 38 women are probably sicker than average
- Prob(death|null hypothesis) =
- Prob(death if treated|treatment matters) =
- Prob(death if untreated|treatment matters) =

Let's look at Colebrook's data on puerperal fever

- Prob(death|null hypothesis) = (42+3)/(210+38)=0.18
- Prob(death if treated|treatment matters) = 3/38
- Prob(death if untreated|treatment matters) = 42/210
- Prob(data|null hypothesis) = $\binom{38}{3}(0.18)^3(0.82)^{35}\binom{210}{42}(0.18)^{42}(0.82)^{168} = 0.00247$
- Prob(new data|drug helps) = $\binom{38}{3}\left(\frac{3}{38}\right)^3\left(1-\frac{3}{38}\right)^{35}\binom{210}{42}\left(\frac{42}{210}\right)^{42}\left(1-\frac{42}{210}\right)^{168} = 0.016$
- G=2ln(0.016/0.00247)=3.74
- *p*=0.053 from G-test with df=1
- Almost statistically significant EVEN THOUGH treated women were likely sicker than average: effect size probably huge

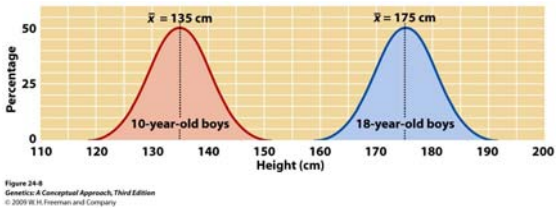
"It behoves us to be very cautious in drawing conclusions as to the curative effect of any remedy upon puerperal infections"

- Why was Colebrook (who did the study) so unsure?
- Even after treating another 64 women (<5% death) then another 100 (8% death) vs. 20% before he started using treatment
  - Maybe the infectious agent had changed
  - Maybe he was seeing healthier women than before
- He didn't use controls because he thought it unethical to deny the drug
- Druin Burch argues that this killed more women by slowing definitive proof of drug effectiveness
- What do you think?

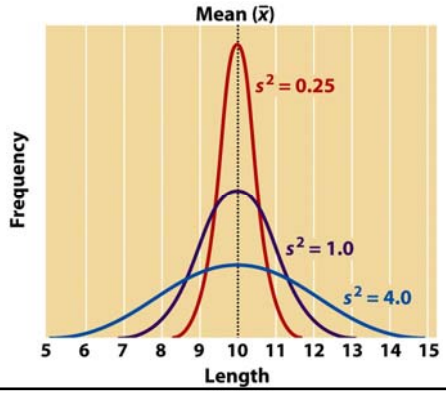
Power analysis

- Last time you rolled dice for patients living vs. dying
- What if data are continuous? Eg,
  - time to death
  - severity of a disease
  - level of a blood marker that correlates with disease
- What are our tools for describing continuous data?
  - Mean
  - Variance
  - Probability distribution

Different mean, same variance



Same mean, different variance



Class quiz

The Belmont Report encompasses three key principles: autonomy, beneficence, and justice. What do each of these mean?

Discussion points from reading

- Exempt vs. expedited vs. Full Board
- We did a class trial on tasting bottled vs. tap water. We will do another one. What are we?
- Apparently, what we are doing is "not research", because our purpose is to contribute to your education, not to generalizable knowledge, so we are none of the above, and don't need to go to the IRB
- I challenge you to try to get me in trouble: come up with a research proposal so good that we end up contributing to generalizable knowledge instead!

If we were doing "real research" rather than a class project, which category would we be?

Exempt or expedited if no more than minimal risk

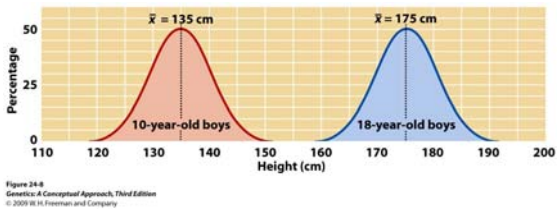
- The probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily life or during the performance of routine physical or psychological examinations or tests.

Exempt vs. expedited

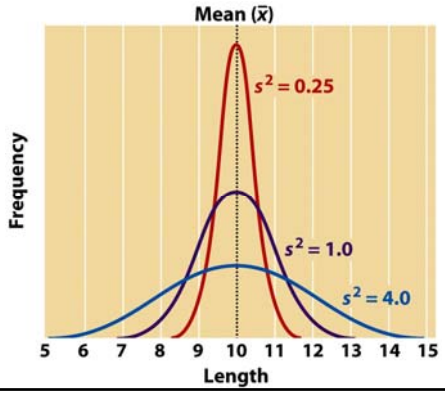
- Exempt (we do a bunch of paperwork to persuade the IRB that they don't have the authority to regulate us) if we study
  - Educational methods
  - Educational tests, eg of cognition, unless confidentiality breach could cause harm
  - Normal food
- Otherwise expedited review
- We should try to provide as much confidentiality as we can as to educational test scores

Back to statistics

Different mean, same variance



Same mean, different variance



Calculate the mean

Data
5
8
2
3
4
7
9
4
3

Calculate the sum

	Data
	5
	8
	2
	3
	4
	7
	9
	4
	3
Sum	45

Mean = sum / number of datapoints

	Data
	5
	8
	2
	3
	4
	7
	9
	4
	3
Sum	45
Mean	5

What if I add a second data set later? What is the mean of all the data?

	Data
	5
	8
	2
	3
	4
	7
	9
	4
	3
Sum	45
Mean	5

	Data
	6
	1
	2
	3
	3
	4
	5
	6
	5
	4
	5
Sum	44
Mean	4

NOT the average of the two means (4.5)
(45+44)/(9+11)=4.45

Calculate the variance and standard deviation

	Data
	5
	8
	2
	3
	4
	7
	9
	4
	3
Sum	45
Mean	5

Start with the deviation of each data point from the mean

	Data	Deviation
	5	
	8	
	2	
	3	
	4	
	7	
	9	
	4	
	3	
Sum	45	
Mean	5	

Start with the deviation of each data point from the mean

	Data	Deviation
	5	0
	8	+3
	2	-3
	3	-2
	4	-1
	7	+2
	9	+4
	4	-1
	3	-2
Sum	45	
Mean	5	

The mean deviation is zero

	Data	Deviation
	5	0
	8	+3
	2	-3
	3	-2
	4	-1
	7	+2
	9	+4
	4	-1
	3	-2
Sum	45	0
Mean	5	0

How big are the deviations, ignoring their direction?

	Data	Deviation	Absolute deviation
	5	0	0
	8	+3	3
	2	-3	3
	3	-2	2
	4	-1	1
	7	+2	2
	9	+4	4
	4	-1	1
	3	-2	2
Sum	45	0	18
Mean	5	0	2

Hocus pocus: use the deviation squared instead

	Data	Deviation	Absolute deviation	Deviation squared
	5	0	0	0
	8	+3	3	9
	2	-3	3	9
	3	-2	2	4
	4	-1	1	1
	7	+2	2	4
	9	+4	4	16
	4	-1	1	1
	3	-2	2	4
Sum	45	0	18	48
Mean	5	0	2	5.33

Mean is dominated by extreme values

The variance is the expected size of a deviation squared

	Data	Deviation	Absolute deviation	Deviation squared
	5	0	0	0
	8	+3	3	9
	2	-3	3	9
	3	-2	2	4
	4	-1	1	1
	7	+2	2	4
	9	+4	4	16
	4	-1	1	1
	3	-2	2	4
Sum	45	0	18	48
Mean	5	0	2	5.33

To calculate variance, divide the sum of dev^2 by n-1 rather than n

	Data	Deviation	Absolute deviation	Deviation squared
	5	0	0	0
	8	+3	3	9
	2	-3	3	9
	3	-2	2	4
	4	-1	1	1
	7	+2	2	4
	9	+4	4	16
	4	-1	1	1
	3	-2	2	4
Sum	45	0	18	48
Mean	5	0	2	48/9=5.33
Variance of the data				48/8 = 6 = 2.45^2

If we knew the true mean, we would divide by n

- We lost a degree of freedom when we used data to calculate the mean (remember G-test)
- Consider the mean of coin-flipping data where heads = 0 and tails = 1. We know the mean is 0.5.

data	1	1	0	0	0	1	1	1	1	1	1	0	0	1	0	Observed mean 10/16=0.625
Absolute deviations based on true mean	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	0.5	True Variance 16×0.5^2/16=0.25
Deviations^2 based on calculated mean	.14	.14	.39	.39	.39	.14	.14	.14	.14	.14	.14	.14	.39	.39	.14	.39	Sum 3.1875 Divide by 16 → 0.2 Divide by 15 → 0.21

Estimated mean and variance in equation form

mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

If the data measures concentration, then variance has units of concentration squared

Mean and deviations have same units as the original data

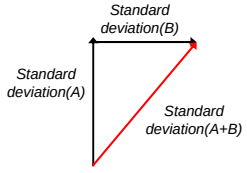
Reverse the effect of taking the square of each deviation by taking the square root at the end

$$\text{variance } s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\text{standard deviation } s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Why take the average of the deviation squared, rather than the average absolute deviation?

Because Var(A+B)=Var(A)+Var(B)



Like Pythagorus' theorem

If A and B are independent, we have a right angle.

Roll 10 times. Calculate the mean, variance and standard deviation of those 10 numbers

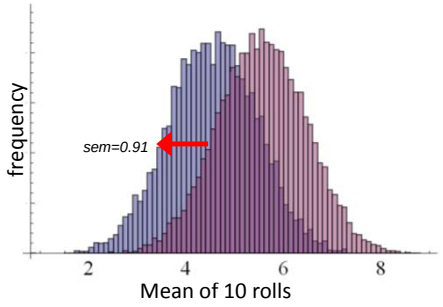
- Left side of room: your dice sample numbers 0-9
- Right side of room: your dice sample numbers 1-10

I.e. depending on where you sit, "0" is either 0 or 10.

Your dice sample numbers 0-9

- Who got mean <2, 2-3, 3-4, 4-5, 5-6, 6-7, >7?
  (the "true" answers are 4.5 and 5.5)
- The "true" standard deviation is 2.87228, you all should have got some value similarly scattered around this number
- What is the standard deviation of the calculated mean?
  (the "true" answer is 0.908295)
  Terminology: standard error of the mean (s.e.m)

Power related to lack of overlap between the 2 curves



Power related to lack of overlap between the 2 curves

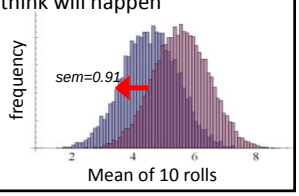
How do we get more power?

What happens to the two curves if

- the effect size is doubled, so 0-9 vs. 2-11?
- everybody rolls 20 times instead of 10 times?

Draw a sketch of what you think will happen

Which will increase power the most?



Class quiz

What is informed consent?

Enough (comprehensible) information

- *Purpose* of the research
- *Procedures* involved in the research
- *Alternatives* to participation
- All *foreseeable risks and discomforts* to the subject (including possible psychological, social, or economic harm, discomfort, or inconvenience)
- *Benefits* of the society and possibly to the subject
- *Length of time* the subject is expected to participate
- *Payment* for participation (if applicable)
- *Person to contact*
- Statement that *participation is voluntary* and that refusal to participate will not result in any consequences
- Subjects' *right to confidentiality and right to withdraw* from the study at any time without any consequences

Power related to lack of overlap between the 2 curves

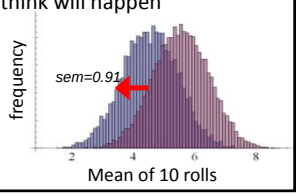
How do we get more power?

What happens to the two curves if

- the effect size is doubled, so 0-9 vs. 2-11?
- everybody rolls 20 times instead of 10 times?

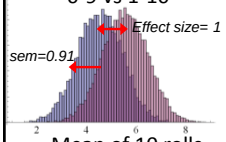
Draw a sketch of what you think will happen

Which will increase power the most?

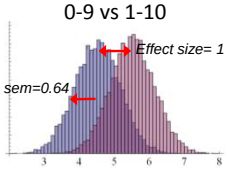


Doubling the effect size makes a bigger difference than doubling the number of dice rolls

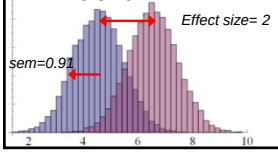
Mean of 10 rolls, 0-9 vs 1-10



Mean of 20 rolls, 0-9 vs 1-10

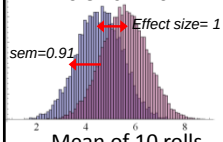


Mean of 10 rolls, 0-9 vs 2-11

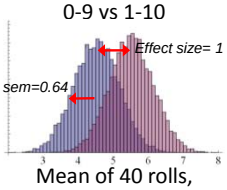


Twice the average difference between groups ≡ four times the number of data points

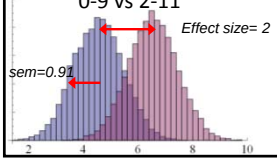
Mean of 10 rolls, 0-9 vs 1-10



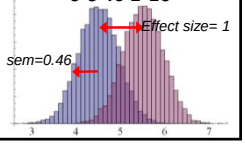
Mean of 20 rolls, 0-9 vs 1-10



Mean of 10 rolls, 0-9 vs 2-11



Mean of 40 rolls, 0-9 vs 1-10

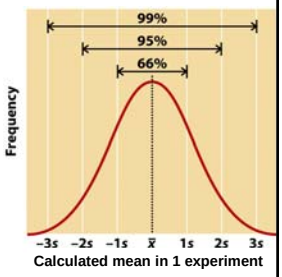


Power (lack of overlap) demands a small standard error of the mean (sem)

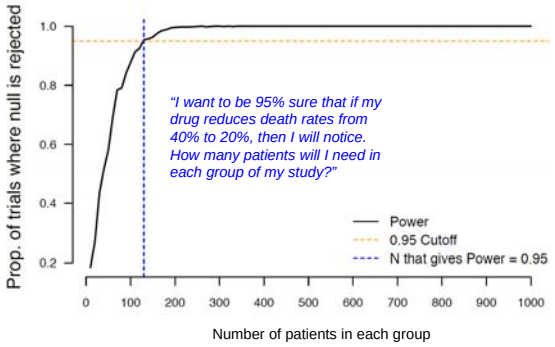
- Mean of patients = sum/n
- Variance of sum = n × variance of 1 patient
- s.d. of sum $= \sqrt{n \times \text{var}_{\text{patient}}}$
- s.e. of mean $= \sqrt{\text{var}_{\text{patient}} / n}$

Power (lack of overlap) depends on the effect size relative to the standard error of the mean

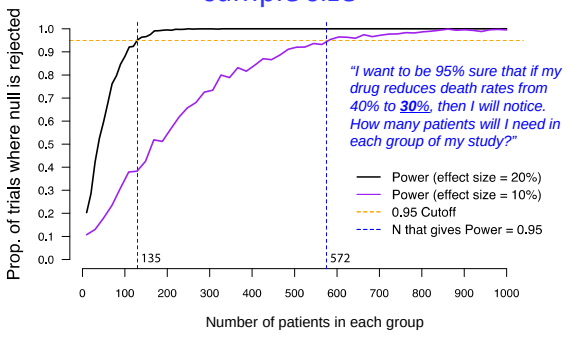
- s.e. of mean = $\sqrt{var_{patient} / n}$
- To detect effects half the size, you need to halve the s.e.m. For that, you need four times as many patients



Power depends on the number of patients

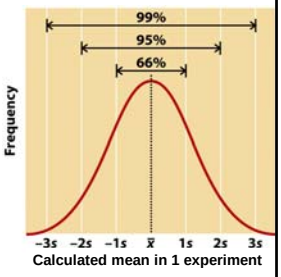


Power is a balance between effect size and sample size



Power (lack of overlap) depends on the effect size relative to the standard error of the mean

- s.e. of mean = $\sqrt{var_{patient} / n}$
- Variance among patients matters too: controlling things more can reduce this
- One aim of a pilot study is to estimate $var_{patient}$ and so calculate how large the final study needs to be to have adequate power.



- How large an effect do I want power to detect *relative to* standard deviation(patients)? This determines *n*.

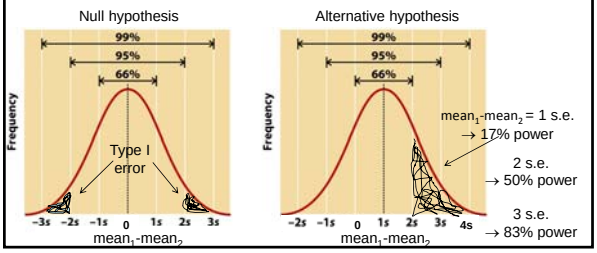
Play with this

http://www.amstat.org/publications/jse/v10n3/aberson/power_applet.html

(To use this, you may need to find the Java program on your computer and create a security exception for this website)

Another way to visualize power as a function of sem using just one curve

- Are 2 groups different? Null hypothesis = no
- Under null, expect $mean_1$-$mean_2$ = 0
- s.e.($mean_1$-$mean_2$) = $\sqrt{var_{patient} / n}$



Which of the following is more likely?

A) I roll 10 dice, and get 7 even numbers
B) 100 people each roll one die, and 70 of them get even numbers
C) equally likely
D) impossible to say

Class quiz

What was the key difference between Fleming's "mould juice" and antiseptics? Why could mould juice work on deep wounds that antiseptics made worse?

Discussion points from reading

- What is your verdict on Fleming and Paine's failure to develop penicillin earlier?
- Why do people associate the discovery of penicillin with Alexander Fleming rather than Howard Florey?
  - Personality, willingness to talk to the press
  - Good connections
  - Branding (Florey did not give a novel name to the active ingredient)

Discussion points from reading

- Waksman vs. Schatz: who deserves credit for streptomycin?
- Work itself was pretty tedious, does the credit go to the design from the top?
- Should this sort of research be done by PhD students / scientists, or by lab technicians?

Discussion points from reading

Sanocrysin randomization study design

- 24 patients put into 2 groups matched as closely as possible
- A single coin flipped for which group got what
- Care taken in blinding

Randomizing patients

- Each pair of students gets a pack of cards to use as "patients".
- Randomize your patients into 2 groups: one group will get the drug, the other the placebo
- I have dice if you want to use them

Randomization methods

- Toss a coin/die for each patient
  - Groups will not be the same size, this reduces power
  - Could make the last few patients balance out group sizes
- Alternate patients
  - But doctors may be able to predict and cheat
  - Put allocations in sealed envelopes (but doctors have been known to hold them to the light)
  - Double-blind (neither doctor nor patient knows what patient is getting)
  - Essential to blind anyone providing a subjective measure of outcome

What if the four suits are four slightly different cancers, and diamonds are the only ones to benefit from this treatment?

- How many diamonds do you have in each group?
- This is similar to a binomial distribution, 52 patients, probability of 0.25 per patient
- What if the number tells you how advanced the cancer's spread is?
- Try randomizing again, this time taking all this into account

Randomization methods

- Randomize only within "blocks" (or even pairs) of similar patients, eg same sex
- Best strategy: control what you can, and randomize whatever is left
- Power goes with s.e.($mean_1$-$mean_2$), so reducing variance in the *difference between the two groups* through matching similar pairs can help

In medicine, what should you control for?

- Sex
- Age
- Disease severity
- Ethnicity
- Pregnancy
- In an education study, we might want to control for
  - instructor
  - time of day

Start thinking about proposal

- Needs to be something ethical and achievable for us to do as a group
- Bring your ideas next lecture to run by your classmates, and refine them in conversation
- Then write them up yourself, due Mar 13. It will be redistributed to other students, with the following assignment being to provide written comments on each other's ideas
  - Study population
  - Intervention
  - Outcome measurement

Instructor's note

I went to the board to draw out

Parallel	Crossover	N of 1
T	TP	TTPPPTTPTPTTPPPPTTPTPT
P	PT	
P	PT	
P	PT	
T	TP	

Class quiz

Why did Cochrane think his "first, worst and most successful" trial was not good?

Study design

Parallel group

- patients randomized into 2 or more groups
- feasible for most tests on individual outcomes
- some other designs get better power

Crossover study design

Receive one treatment then the other (in random order), so that each patient acts as their own control

- Detectable effect size goes with $\sqrt{\text{var}_{\text{data}}/n}$
  - In parallel groups this means variance between patients
  - With a crossover design it means variance in the amount a patient changes between treatments
- Works only for chronic not acute conditions
- Make sure one treatment has worn off before you start the next. If one treatment won't wear off, it has to come last; inability to do both orders is a problem

If disease is slow and treatment fast, you can do an "N of 1" study, where ALL data points come from the same patient

More study designs

1. Split-body (right vs. left sides)
2. Cluster (groups rather than individuals are assigned to a treatment
   - eg sections of a class rather than individual students
   - unavoidable where you can't target an individual without affecting the whole group
   - hard to get enough independent data points
3. Factorial (multiple treatments studied at once, patients receive different combinations)
   - First randomization was by Fisher 1923, in agriculture. He divided the land up into small plots and assigned fertilizer, water etc. to each plot in a factorial way

Which study design would you choose to test the merits of a…

- pill for acute back pain
- same pill but for chronic back pain
- surgical checklist
- acne cream
- traffic light timing scheme for Tucson

Will it work in the real world?

- Efficacy trials are in strict, ideal conditions with closely vetted patients
- Effectiveness trials are in more usual medical settings
- Efficacy studies are normally done first
  - low $var_{patient}$ to get higher power
  - less relevant to "real life"

After you randomize…

- Drug vs. placebo may change the dropout rate
- This can change the results
- Focus on intent to treat, then try to follow up all outcomes, even after dropout
- Don't restrict analysis to the patients that comply; follow up EVERYTHING that happens after randomization, this is the statistically sound option
- Intent-to-treat analyses focus on effectiveness, not efficacy
- If your hypothesis is about efficacy only, don't do intent-to-treat, but be aware that results are subject to bias

Homework

- Homework 4 on randomization and study design is now available and due in one week on Mar 13.
- You also need to write a research proposal, ALSO due Mar 13
  - Think about study design, eg
    - crossover trials
    - 2 groups vs. more than 2 groups
  - If appropriate, identify and address any aspect(s) of the study that will be particularly challenging
  - Come to office hours to discuss your ideas!

What some past classes did

- Do workout supplements improve bench-pressing ability?
- Do email reminders increase the proportion of Gen Chem students completing their homework on time?
- Does texting people who signed up for a blood drive increase the rate of people showing up?
- Does toothpaste help scrub coffee stains off cow teeth?
- Do men vs. women open doors more often for men vs. women?

"2-answer" formats can work well

Eg, ask people in a science class whether they find a particular argument persuasive, testing whether it makes a difference if the argument is presented with vs. without a table of supporting data.

Proposal format

Every proposal is different, with different emphases, but all must address, to the extent necessary,

- motivation for the proposal question
- starting population
- intervention
- study design (and why that design was chosen)
- outcome measure
- ethics

Break up into groups of 2 or 3

- Pitch your proposal idea to your neighbor
- Make helpful suggestions to improve them
  - What are the practical details?
    - Where do you get study subjects?
    - Intervention
    - Study design
    - Outcome measurement
  - Is it doable?
  - Is it ethical?
  - Are there ways of improving the design?

Class quiz

What allowed the British Medical Research Council (MRC) to perform a randomized trial of the effects of streptomycin on pulmonary tuberculosis? In particular, how did they persuade presumably reluctant doctors to include a hospitalization-only control group?

Discussion points from reading

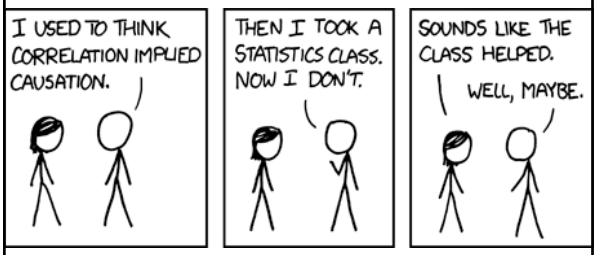
- First statistically analyzed randomized trials were in agriculture, done by Fisher in 1920s
- In 1931, Sanocrysin for tuberculosis was a not-quite-randomized controlled trial 24 patients, placebo did better, no formal statistics
- Colebrook historical data on puerperal fever published 1936, Fisher's statistical methods were not used
- 1948 streptomycin for tuberculosis, 109 patients, no placebo because injections were so painful, drug did better ($p<0.01$), but only at first. Patients were not told about trial.
- Delay partly because statistics was new, partly because of doctors' resistance

Discussion points from reading

Correlation vs. causation

- Galton noted that kings had more people praying for them than other aristocrats, but lived less time. Does prayer shorten life? Or are kings under so much stress that they would die even earlier if not for all the prayers?

Observing what happens naturally (correlation) is not enough. To measure causation, you need an intervention and a control group



Correlation vs. causation

- Your proposal should include a controlled intervention.
- Fisher took this to extremes. People who smoke get lung cancer, but this is not a randomized controlled trial. Might there be a gene that causes both smoking and lung cancer?

Homework reminders

- Bring FOUR hard copies of your research proposal with you to class on Thursday
- ALSO upload research proposal to D2L Dropbox before class
- Bring ONE hard copy of homework 4 to class on Thursday

Clinical trials need

1. a starting population
2. an intervention
3. an outcome measure of endpoints

Possible endpoints for malaria

- Feeling better soon (reduced fever)
- Time to recovery from this episode
- Death during this episode
- Whether disease recurs
- Whether/when disease kills you in the end

Reductionism

Science works by making things simpler
(perhaps too simple, according to reading)

1. Endpoint
2. Disease cause

Assume disease has a single cause

- For an infection, the cause is
  - the pathogen
- For cancer
  - the tumor
- For gastrointestinal bleeding
  - the bleeding vessel or ulcer
- Genetic diseases
  - one gene, one mutation, one disease
- Target treatment to that one cause

Do diseases have just one cause?

- The disease, and not the person affected by it, becomes the central focus.
- A young immuno-compromised man with pneumococcal pneumonia usually gets the same antibiotic treatment as an elderly woman with the same infection.
- For other diseases (eg cancer), we may be lumping together cases that are actually quite different
- Aim is to identify diagnoses that are truly the same, and determine the best treatment for each diagnosis (not for each patient)

Reductionism

1. Endpoint
2. Disease cause
3. Treatment
   - Polypharmacy (i.e. complex potions with esoteric rare ingredients prepared a particular way) vs. single active ingredient

Why isolate quinine rather than use cinchona bark?

- More reliable doses
- Might taste better or have fewer side effects from other components of the same plant
- Learn to synthesize it
  - easier than accessing the trees
  - lower cost
- More “scientific”: if you know the chemical, you can understand HOW it works
- Note that no chemistry will tell you WHETHER it works: only a clinical trial can do that

If something is too high/low, treatment tries to bring it back

- What if the "pathology" we see is actually the body's way of compensating for something else?
- Clinical trials may show that
  - more calcium does not help low calcium levels
  - high blood pressure means high risk, but drugs that lower blood pressure do not necessarily reduce that risk

Are we even measuring right?

- Is it the "normal" range of some indicator that matters?
  - or how it responds to stress (eg insulin in response to fasting vs eating sugar)
  - or how it oscillates (eg circadian rhythm)
  - or more complex patterns (eg with heart rate irregularities and variability)

Reductionism

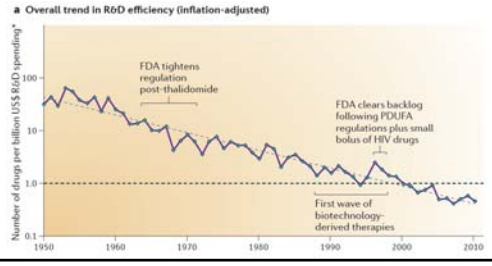
1. Endpoint
2. Disease cause
3. Treatment
   - Single active ingredient
   - Homeostasis model of disease
     (find out what is not normal and bring it back)
4. Disease mechanism
   - Much of modern biomedical "pure" science

Ideal reductionist pipeline

1. Identify patients with same disease
2. Which organs are affected?
3. Which cells in those organs?
4. Which molecules in those cells?
5. Which molecule needs to change to fix things?
6. Design a drug to make that molecule change
7. Test it on the purified molecule
8. Test it on cell culture
9. Test it on animal model of disease
10. Test it on patients
11. Cure!

Reductionism

- Divide and conquer: understand a system by understanding its parts
- One model of how science progresses, but something is going wrong



Where did good drugs, like aspirin and antibiotics, come from?

- Trial and error of more or less random things
  - willow bark
  - a dye molecule (whose dye portion was irrelevant)
- Drug to treat restricted blood flow to the heart didn't work, but had some strange side effects
  - now sold as Viagra

Appeal of reductionism

- Understanding obscure molecular pathways makes us possessors of rare knowledge and increases the status of doctors and biomedical researchers
- Scientific establishment works wells with division of labor, everybody can achieve something with a divide and conquer strategy

Pharmaceutical Industry Historical Reality

- Take lots of random drugs and throw them at cell culture and/or animal model
- Many drugs that work in culture don't work in animals, the drug "pipeline" is broken
- Atoxyl against trypanosomes worked on live animals but NOT in culture
- Still too toxic: drugs need to kill the pathogen, but not the patient

Instead of basic science → drug, accidental drug success can shape our basic science beliefs

- In 1950s, tried to treat schizophrenics with a drug that tweaked the balance of the brain's neurotransmitters
- This sent patients into bouts of euphoria, making things worse
- Euphoria seems like a good idea to treat depression
  - imipramine, the first tricyclic, marketed as Tofranil in 1958
  - Similarly, 1st MAOI drug was tried for TB, made even terminally ill patients more cheerful, optimistic, and physically active
- Serious side effects. Tricyclics blanket a broad range of brain chemicals; new "selective" serotonin reuptake inhibitors or SSRIs (eg Prozac) zeroed in on one: serotonin
- No more effective, but fewer harms and side effects like sedation

We have no *direct* evidence for serotonin or other "chemical imbalance" in the brains of depressed patients

- Logic is that because SSRIs work, depression must be a serotonin imbalance
- By the same logic
  - pain is a chemical imbalance of opiates
  - fever is a chemical deficiency of aspirin
- Later this course, we will question whether SSRIs work
- Tianeptine, a monoamine/serotonin reuptake *enhancer*, is also licensed as an antidepressant

Class quiz

If a drug successfully improves the symptoms of genetically modified knockout mice, why might it NOT work on humans with similar symptoms?

Make sure you leave here with 3 different proposals, none from students you have already discussed ideas with

- Rank the three proposals as to which one you want the class to do (1 is best, 3 is worst). Rank will NOT affect the other students' grades
- Comment (including suggested improvements) on
  1. feasibility
  2. ethics
  3. interest
- You will be graded on the quality of your comments.
- Upload comments to dropbox, clearly indicating which is your name, and which is the name of the proposal's author. Start a new page for each of the 3 proposals you review.

We will pick a shortlist

- You will get an email during Spring Break naming the finalists
- Read those proposals (they will be posted under D2L/Content) before the first class after spring Break, and arrive to class ready to discuss them and vote.

Reductionism

1. Endpoint
2. Disease cause
3. Treatment
   - Single active ingredient
   - Homeostasis model of disease
     (find out what is not normal and bring it back)
4. Disease mechanism
   - Much of modern biomedical "pure" science
   - Both your reading and I are skeptical on this one

Basic research that made a big difference to medicine I

- Germ theory
  - Hygiene
    - Although Semelweiss was able to prove the value of handwashing without knowing about germ theory
  - Isolating particular infectious agents to use for vaccine production and to screen drugs against
- Basic genetics underlies screening for treatable genetic diseases such as phenylketonuria

Basic research that made a big difference to medicine II

Animal / cell models of disease allow drugs to be screened (with or without a hypothesis about which drugs are likely to help)

- Eg syphilis cure came from screening lots of compounds against first an ape disease model, then more rapidly a rabbit. Compound 606 worked.
- Not all animal models are close enough to human to be useful

Basic research that made a big difference to medicine III

- Some molecular treatments were driven by mechanistic understanding
  - hormone therapies, eg insulin, contraceptive pill
  - vitamin deficiencies, eg scurvy, rickets
  - some antiviral drugs
  - mimics of existing drugs, hopefully more specific / less allergenic
- Imaging technologies, eg X-rays, or EEG/ECG
- Tissue culturing technology allowed polio virus to be produced in bulk, leading eventually to a vaccine

Translational research

- Will somebody please take the masses of reductionist biology being done and translate it into something that might help patients?
- Areas of as-yet-unfulfilled promise
  - Stem cells for regeneration
  - Gene therapy
- How well do you need to understand a disease in order to cure it?

Placebo effects

History of medicine mostly
= history of placebos

Multiple ways to use the word "placebo"

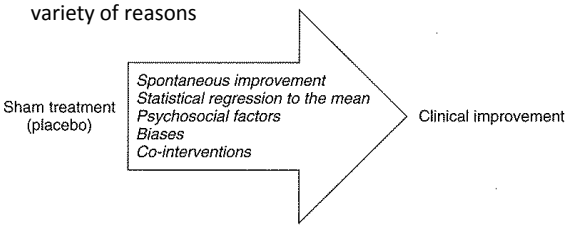
1. Placebo is the sham treatment itself
   - Sugar pill
   - "Sham" surgery (general anesthetic, open up patient, then sew them back up)
   - "Sham" acupuncture (wrong position or toothpicks not needles)
   - Sometimes it is hard to design a placebo, eg for psychoanalysis

Multiple ways to use the word "placebo"

1. Placebo is the sham treatment itself
2. "Placebo group" = control group
   - Placebo / control group treated ALMOST the same way as treated group
3. "Placebo effects" are anything that happens in the placebo group
4. I will use "placebo response" to mean that giving the sham treatment genuinely causes an improvement for the patient

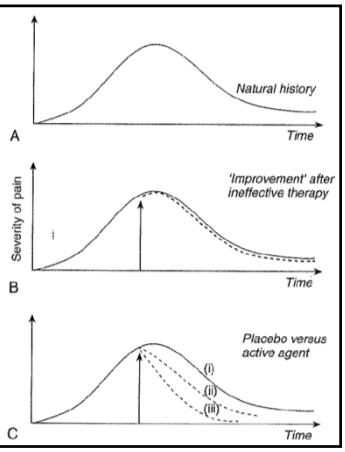
Placebo effects

- Disease may improve even in the control group, for a variety of reasons



- Psychosocial factors / mind over matter / "true" placebo response is only one of these

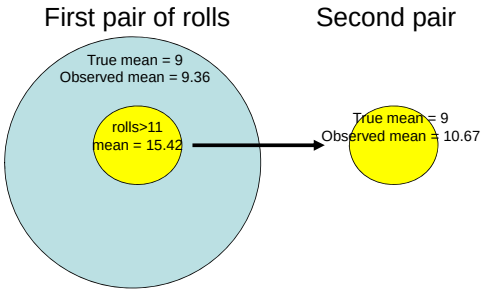
Spontaneous improvement



Who has "lucky hands"?

- Roll the dice twice and sum them up
  - class mean?
- If you got over 11, again roll twice and take the sum
  - mean of those with lucky hands?
- Lucky hands don't exist: instead we have regression of the lucky back down to the class mean

Regression to the mean: no "hot hands"



Regression to the mean

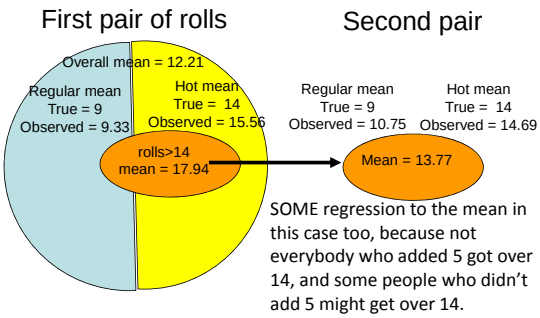
- Imagine patients whose symptoms fluctuate from 0-1
- Clinical trials usually enrol extreme cases
- Study enrols if >0.7

Random numbers	
Series 1	Series 2
0.71	0.21
0.11	0.31
0.70	0.74
0.23	0.28
0.49	0.12
0.26	0.83
0.35	0.18
0.63	0.15
0.25	0.90
0.55	0.34
0.95	0.64
0.09	0.57
0.92	0.55
0.32	0.11
0.86	0.32
0.59	0.03
0.02	0.63
0.96	0.18
0.36	0.57
0.68	0.30
0.04	0.51
0.63	0.83
0.80	0.34
0.05	0.54
0.55	0.57
0.93	0.52
0.81	0.61
0.72	0.95
0.24	0.25
0.14	0.51
m=0.49	m=0.45

What if there really are "lucky hands"?

- Roll twice more, and half of you add 5 to the sum.
  - Class mean? 12.21
  - Regular mean? 9.33
  - Lucky mean? 15.56
- Roll again if you get over 14
  - Mean of the lucky 14.69
  - regular 10.75
  - overall 13.77

Regression to the mean: +5 "hot hands"



SOME regression to the mean in this case too, because not everybody who added 5 got over 14, and some people who didn't add 5 might get over 14.

Class quiz

How did Archie Cochrane test whether cardiologists knew how to interpret electrocardiograms (or alternatively, whether dentists knew how to diagnose teeth)? What did he find?

Discussion points from reading

- Randomized controlled trial not only of drugs, but of bedrest as a treatment for TB
- Do you agree with Crofton (and with Druin Burch) that randomized controlled trials are boring? Cochrane thought they weren't.
- Cochrane wanted to apply the method to education, criminal justice etc. – just like this class!
- Can you imagine being fooled by Cochrane's bait-and-switch presentation of results?

Let's choose a project!

- Feasibility
  - How hard is it to do?
  - Will we have enough power to get an answer?
- Ethics
- Interest
  - Do we already know what the answer is?
  - Do we care?
  - Is it fun?

Grades for project

- 5% of your grade depends on your participation in a class project. We will pick 5 projects, enough to give all of you a role.
- This participation can and should take many forms, not necessarily or only being an experimental subject
- Everybody must do something (or else get 0/5), better to volunteer repeatedly!
- Those whose proposals are chosen have a head start.

Let's choose projects!

1. Biotin speeds up hair and nail growth
2. Music genre affects blood pressure, heart rate, and subjective emotional intensity
3. Distracting hand decorations lower performance on an arithmetic exam
4. Men benchpress more with a woman sitting on their lap
5. Dynamic vs. static stretching for short-term flexibility gain
6. Different stain removers
7. Nike logo makes people want to buy a jacket
8. Coffee stimulates defecation in Efe's friend (N of 1)
9. Coffee causes flatulence

The winners are…

1. Music genre affects blood pressure, heart rate, and subjective emotional intensity
2. Men benchpress more with a woman sitting on their lap
3. Dynamic vs. static stretching for short-term flexibility gain
4. Different stain removers
5. Nike logo makes people want to buy a jacket

Class quiz

Thalidomide had no obvious side effects for the vast majority of people who took it. Why were the dangers of thalidomide recognized as quickly as they were?

Class projects

1. Music genre affects blood pressure, heart rate, and subjective emotional intensity
2. Men benchpress more with a woman sitting on their hips
3. Dynamic vs. static stretching for short-term flexibility gain
4. Different stain removers
5. Nike logo makes people want to buy a jacket

Tasks to allocate

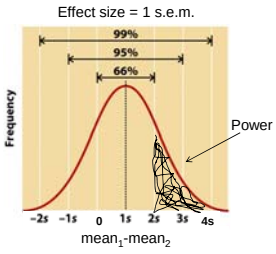
- Press officer to promote our work on all projects (Zach and Austin)
- Within each group
  - Project leader: the buck stops with you
  - Details of study design and implementation
    - maybe different people volunteer for different aspects, eg choosing music and photos vs. measuring heart rate and blood pressure
  - Informed consent forms, if needed
    - not for the stains!
    - may need to be a more minimal protocol for the Nike jacket study. Study involves deception, so we need to DEBRIEF participants afterwards. And maintain their anonymity.
  - Co-ordinating recruiting of subjects
  - Randomization, recording and compiling data:
    - especially important with blinding
  - Data analysis / statistics (Parris and I will help)
  - Write a report
- Participating in the studies / recruiting more subjects

Pilot study needed?

- To test that the protocol works
- To figure out how much natural variation there is, and so how large a study will be needed to detect a given effect size
  - in practice, most class projects tend to be pilot studies on this count

Power

- How large do we expect the effect size to be?
- How does this compare to the standard error of our estimate?
  - For parallel groups, s.e. is related to $\sqrt{\text{var}/n}$ where var = variance between subjects
  - For crossover, var = how variable is the difference between 2 measurements of the same subject



Go to your groups

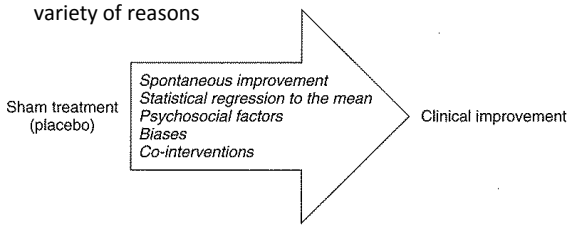
1. Music genre affects blood pressure, heart rate, and subjective emotional intensity
2. Men benchpress more with a woman sitting on their hips
3. Dynamic vs. static stretching for short-term flexibility gain
4. Different stain removers
5. Nike logo makes people want to buy a jacket

Start to refine the study design and allocate roles (continue outside class)

- Press officer to promote our work on all projects (Zach and Austin)
- Within each group
  - Project leader: the buck stops with you
  - Details of study design and implementation
    - maybe different people volunteer for different aspects, eg choosing music and photos vs. measuring heart rate and blood pressure
  - Informed consent forms, if needed
    - not for the stains!
    - may need to be a more minimal protocol for the Nike jacket study. Study involves deception, so we need to DEBRIEF participants afterwards. And maintain their anonymity.
  - Co-ordinating recruiting of subjects
  - Randomization, recording and compiling data:
    - especially important with blinding
  - Data analysis / statistics (Parris and I will help)
  - Write a report
- Participating in the studies / recruiting more subjects

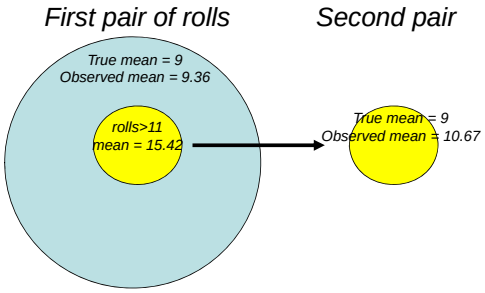
Placebo effects

- Disease may improve even in the control group, for a variety of reasons



- Psychosocial factors / mind over matter / "true" placebo response is only one of these

Regression to the mean: no "hot hands"

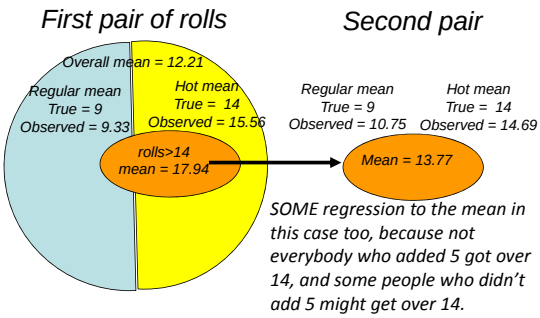


Regression to the mean

- Imagine patients whose symptoms fluctuate from 0-1
- Clinical trials usually enrol extreme cases
- Study enrols if >0.7

Random numbers		Selection ≥ 0.70	
Series 1	Series 2	Series 1	Series 2
0.71	0.21	**0.71**	0.21
0.11	0.31	—	—
0.70	0.74	**0.70**	0.74
0.23	0.28	—	—
0.49	0.12	—	—
0.26	0.63	—	—
0.35	0.18	—	—
0.63	0.15	—	—
0.25	0.90	—	—
0.55	0.34	—	—
0.95	0.84	**0.95**	0.84
0.09	0.57	—	—
0.92	0.55	**0.92**	0.55
0.32	0.11	—	—
0.86	0.32	**0.86**	0.32
0.59	0.03	—	—
0.02	0.63	—	—
0.96	0.18	**0.96**	0.18
0.36	0.57	—	—
0.68	0.30	—	—
0.04	0.51	—	—
0.63	0.83	—	—
0.80	0.34	**0.80**	0.34
0.05	0.54	—	—
0.55	0.57	—	—
0.93	0.52	**0.93**	0.52
0.61	0.61	—	—
0.72	0.95	**0.72**	0.95
0.24	0.25	—	—
0.14	0.51	—	—
m=0.49	*m*=0.45	*m*=0.84	*m*=0.52

Regression to the mean with +5 "hot hands"



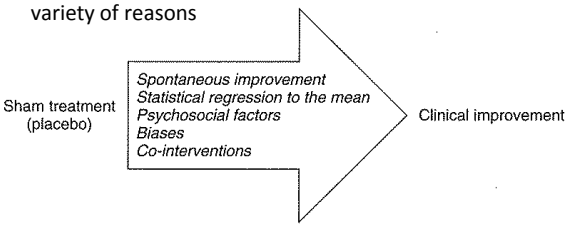
SOME regression to the mean in this case too, because not everybody who added 5 got over 14, and some people who didn't add 5 might get over 14.

Can you see regression to the mean in your daily life?

- You try out a new restaurant and you LOVE it, best thing ever.
- A few months later, you finally make it back there and it isn't nearly as good.
- Has the restaurant got worse?

Placebo effects

- Disease may improve even in the control group, for a variety of reasons



- Psychosocial factors / mind over matter / "true" placebo response is only one of these

Biased reporting

- Patients may exaggerate their initial symptoms if they want to be part of a trial
- Patients later want to please the doctor, who spent time and effort trying to help
- Unblinding the doctors may reduce placebo effects if doctors then
  - judge patient improvement differently
  - give different cues to the patient, affecting both their true placebo response, and any bias in their reporting

Only a drug helped asthmatics breathe

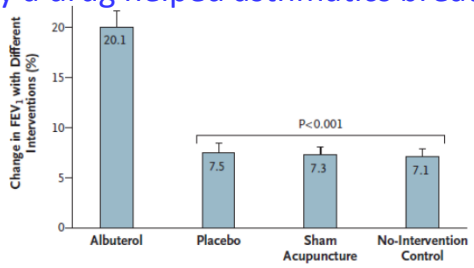


Figure 3. Percent Change in Maximum Forced Expiratory Volume in 1 Second (FEV_1) with Each of the Four Interventions.

The relative improvement in FEV_1 achieved with albuterol was significantly greater than that achieved with each of the other three interventions (P<0.001). No other differences among the four experimental conditions were significant. T bars indicate standard errors.

But patients reported feeling better with placebo

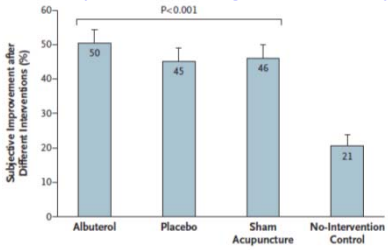


Figure 4. Percent Change in Subjective Improvement with Each of the Four Interventions.

The relative improvement in subjective outcomes, assessed with the use of a visual-analogue scale (with 0 indicating no improvement and 10 indicating complete improvement), was significantly greater with the albuterol inhaler, placebo inhaler, and sham acupuncture interventions than with the no-intervention control (P<0.001). No other differences among the four experimental conditions were significant. T bars indicate standard errors.

More biases

- Scaling bias: "on a scale of 1-5 where 1 is no improvement and 5 is complete recovery…"
  - What if you get worse?
- Signal detection ambiguity (less likely to report unexpected deterioration than expected improvement when either is a close call)

Co-interventions may be the cause of improvement

- eg cough syrup where the syrup itself, not the “active” ingredient, turns out to reduce symptoms

Class quiz

Thalidomide has been approved by the FDA to treat which disease? In contrast, which disease is it usually prescribed for? Why didn't drug companies ask the FDA to approve thalidomide for its most common use, rather than a rare use?

Discussion points from reading

- In 2006, the FDA approved thalidomide (in combination with dexamethasone) to treat multiple myeloma. Derivative drugs approved since then.

Discussion points from reading

- Comments on the ethics of regulating drug trials?
- Druin Burch feels
  - more should be done to make quality trials happen
  - less done to protect trial participants if need be (eg less extensive informed consent, less free access to quality treatment after trial ends)
- Ethics of experimenting on people systematically vs. ethics of widespread use of an untested treatment

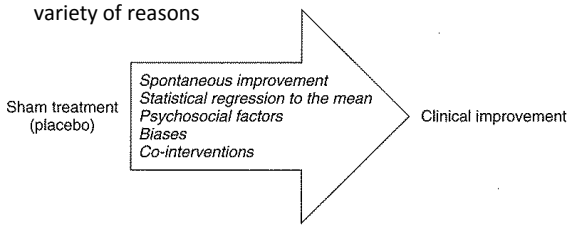
Any updates on class projects?

1. Music genre affects blood pressure, heart rate, and subjective emotional intensity
2. Men benchpress more with a woman sitting on their hips
3. Dynamic vs. static stretching for short-term flexibility gain
4. Different stain removers
5. Nike logo makes people want to buy a jacket

Groups 3,5 (and 1 a bit too), good to see you making progress via D2L Discussions!

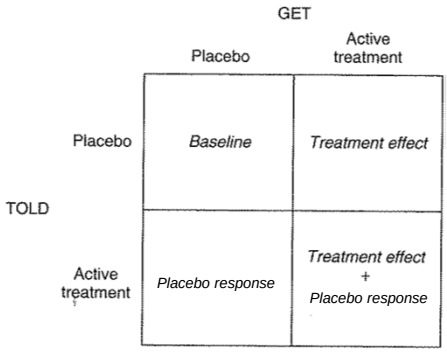
Placebo effects

- Disease may improve even in the control group, for a variety of reasons



- Psychosocial factors / mind over matter / "true" placebo response is only one of these

How to study the "true" placebo response: deceive the patients



Is the placebo response real?

- Depends on the disease
- The best (but not always ideal) evidence is for diseases with strong neuro-psychological basis
  - Very best evidence is for pain

Placebo for migraine

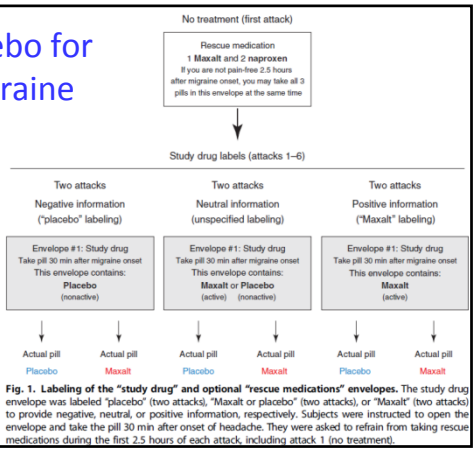


Fig. 1. Labeling of the "study drug" and optional "rescue medications" envelopes. The study drug envelope was labeled "placebo" (two attacks), "Maxalt or placebo" (two attacks), or "Maxalt" (two attacks) to provide negative, neutral, or positive information, respectively. Subjects were instructed to open the envelope and take the pill 30 min after onset of headache. They were asked to refrain from taking rescue medications during the first 2.5 hours of each attack, including attack 1 (no treatment).

Both drug and labelling matter

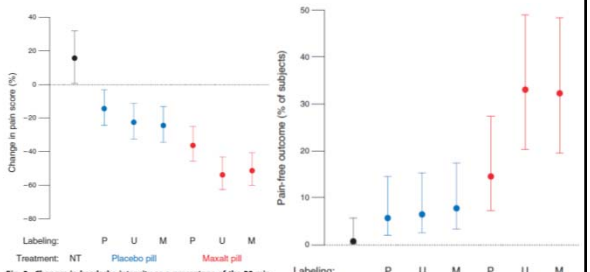
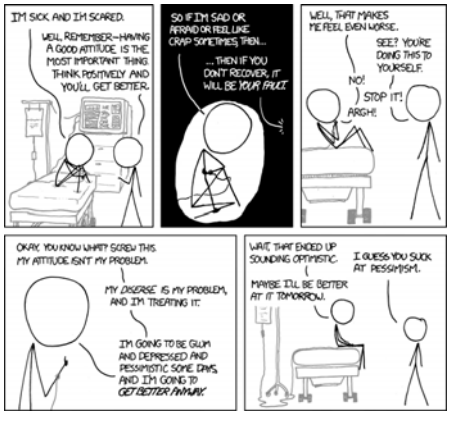


Fig. 3. Changes in headache intensity as a percentage of the 30-min pain score. The data are estimates for the seven experimental conditions, with 95% CIs, from the generalized linear mixed model (table S8). The estimates for the three types of information (labeling) are grouped according to whether the treatment was a placebo pill (blue) or a Maxalt pill (red). The within-subjects design of this study allowed each subject to serve as his or her own control, which substantially increased statistical power. Consequently, 95% CIs cannot be interpreted in the same manner as in a typical between-subjects study. Thus, two groups can differ significantly even when the mean for one group falls within the 95% CI for the other group. NT, no treatment; P, "placebo" label; U, unspecified "Maxalt or placebo" la-

Fig. 4. Percentage of subjects who reported being pain-free 2.5 hours after onset of headache. The data are estimates for the seven experimental conditions, with 95% CIs, from the mixed-effects logistic regression model (table S13). The estimates for the three types of information (labeling) are grouped according to whether the treatment was a placebo pill (blue) or a Maxalt pill (red). NT, no treatment; P, "placebo" label; U, unspecified "Maxalt or placebo" label; M, "Maxalt" label.

Is the placebo response real?

- Depends on the disease
- The best (but not always ideal) evidence is for diseases with strong neuro-psychological basis
  - Very best evidence is for pain
    - Not just a reporting bias: placebo response to pain can be blocked by the opioid antagonist naloxone
- Remember all the other placebo effects before jumping to conclusions about a true placebo response
- There is NO evidence for a placebo response with cancer
  - i.e. positive thinking and hope may make your remaining life more pleasant, but it won't help you live longer



Mechanism of true placebo response

1. Expectations
   - Very subtle differences may be enough to elicit very different placebo responses
   - a nocebo is something that you expect will make things worse, and therefore does
2. Conditioning (eg pill of a certain size and shape worked before)

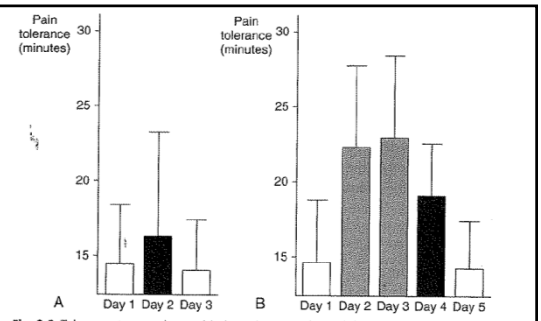
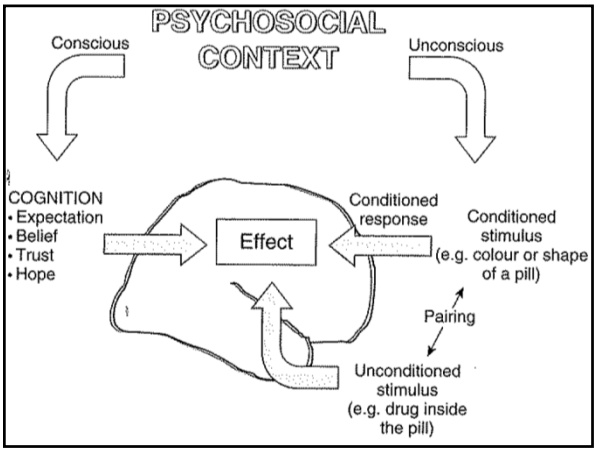


Fig. 2.6 Tolerance to experimental ischaemic arm pain. In (A) a placebo (suggestion of analgesia) is given for the first time on day 2 (days 1 and 3 represent control conditions) as shown by the black column. In (B) the placebo is given on day 4 after administration of the painkiller ketorolac on days 2 and 3 (shaded columns). Days 1 and 5 represent control conditions. It can be seen that the placebo given after pharmacological preconditioning induces greater placebo analgesia than placebo given without preconditioning. From Amanzio and Benedetti 1999 with permission from



SSRI antidepressants

- ~80% of the effect of today's antidepressant drugs comes from placebo effects
  - we don't know how much of that is a true placebo response
  - remainder is 2 points on 62-point Hamilton Depression scale: clinically insignificant
- Trials may not be blind: "this drug is strong enough to make me vomit and hate sex, so it must be strong enough to lift my depression"
  - This might account for the remaining 20%!
- Effect is a bit stronger against severe depression, but this could be due to larger doses with less blinding
- No dose-response curve is worrying

Vertebroplasty

- Osteoporosis can create painful cracks in the spine
- Stick needle into spine and pour cement to fill in the cracks: patients reported dramatic relief
- Procedure became standard
- Eventually, two blinded randomized trials with realistic-seeming SHAM SURGERY
  - Local anesthetic
  - Pressure on back
  - Same chemical smell
- Hard to recruit patients, they didn't want to risk placebo
  - One trial used crossover design: after 1 month, if no relief, they were allowed to have other treatment
- Placebo group experienced just as much relief. Relief in both groups was mostly short-term.

A more powerful belief in the medicine might lead to a stronger placebo response

- Pouring concrete down the spine sounds powerful
- Some studies suggest that expensive placebos work better than cheap ones

Fun video

http://www.youtube.com/watch?v=yfRVCaA5o18

Placebo means "I shall please"

- Many doctors routinely give placebos, eg
  - For inpatients, sugar pill or saline injection instead of a powerful painkiller
  - Vitamin pills (no need to lie: "some people find this helps")
  - Antibiotics
- Some doctors think placebos are ethically fine, others do it but are reluctant to admit it
- What do you think?

Homework 5 is available and due in 1 week

Class quiz

The MRC rejected O'Brien's first application to do a randomized trial on the effects of aspirin, on the grounds that too many patients were needed, making it too expensive. Then they accepted his revised application. Why, what had changed?

Discussion points from reading

- More statistical power in preventative medicine studies if highest risk patients are recruited
  - Balance of benefits vs. harms then unknown for lower risk patients
- Post-heart attack, aspirin increases survival from 88% to 91%

Discussion points from reading

- Doctors weren't satisfied to know that aspirin prevents blood clots: they wanted to know the *mechanism* before they prescribed it
  - reductionism as an end rather than a means
- Drug with obvious side effects (heavy bleeding) looks more powerful
- Different endpoints: death vs 2nd heart attack

Healthcare system

Who controls medical treatments?

- Doctors
- Those offering options to the doctors
  - Pharmaceutical companies
  - Medical device companies
- Those limiting the options
  - Food & Drug Administration (FDA)
  - Patients
- Those educating doctors about which option is better
  - Medical schools
  - Continuing medical education (CME), sometimes funded by pharmaceutical companies
- Cost structure can change the doctors' incentives for different options
  - How much are they reimbursed by insurance companies or Medicare?

How do doctors get paid?

- In the old days, you went to the doctor, you paid
- These days, a few nights in intensive care are so expensive that few people could manage
- Buy insurance to spread the risk
  - Pay every month, no budget shock with unexpected bills
  - Some people are healthy until they drop dead. Others have transplants that cost a fortune for the rest of their life. Everyone pays the same.
- Insurance does not necessarily make health care cheaper
  - Large companies may have more bargaining power over prices, but that may displace costs onto others.
  - Insurance companies take a cut, and may increase paperwork for doctors.

The problem with insurance

- Makes sense not to buy insurance, then rush out and buy it as soon as you get sick
- This is why insurance policies used to exclude "pre-existing conditions"

Different kinds of insurance

1. Buy insurance yourself. Sick (or old) people pay more than healthy ones. If you are sick enough, insurance is unaffordable.
   - When the ER treats you, they charge everyone else more in order to cover that cost.
2. Your employer buys insurance for all employees. Employees don't pay the premium themselves, so they don't opt-out just because they are healthy.
   - Historical loophole to get around post WWII wage controls
   - The price usually depends on the past medical costs of that company's employees. In a "self-funded" program for a large employer, the employer simply pays directly, without using an insurance company.
   - Excludes the self-employed and those without "good" jobs
   - Favored by tax system
3. Buy insurance through your taxes (eg Medicare). Everyone covered.
   - Rich pay more than the poor. Those dying young (often poor) subsidize Medicare

2007 Census Results

- 59.3% employment-based private insurance
- 8.9% bought private insurance themselves
- 13.8% Medicare (government insurance)
- 13.2% Medicaid (social welfare program, arguably not insurance)
- 3.7% military health care
- 15.3% uninsured

(Doesn't add up to 100% because people may have had different forms of cover during the year)

What do other rich countries do?

1. In many, government pays for everyone.
   - Often extra private insurance for perks like a private hospital room
2. In others, insurance is private, but compulsory.
   - Everyone must buy it (the poor get a subsidy or voucher) and insurance companies must offer a minimum coverage at a maximum price to everyone.

- The Patient Protection and Affordable Care Act (Obamacare) did many things, including steps to promote evidence-based medicine. Most controversially, it moves the US closer to #2.

Controversial parts of the Affordable Care Act

- "Individual mandate" for everyone to buy health insurance by March 31, 2014 (i.e. 3 days ago)
  - Fines will start as part of 2014 taxes, paid in Apr 2015
  - Subsidies if you don't earn enough
- Insurers banned from refusing someone or taking into account pre-existing conditions

A few other aspects of the Affordable Health Care Act

- Expand Medicaid to slightly less poor families (some states have refused this Federal money, but Arizona accepted)
- Insurers are required to include adult children up to age 26
- Other things too
- Please test your knowledge at http://healthreform.kff.org/quizzes/health-reform-quiz.aspx

Essay topic

How do you want the US healthcare system to work?

- Pay particular attention to who pays and how.
- If you prefer, you can focus on one aspect of the system rather than the overall design.

How do you want the US healthcare system to work?

- Guideline: 750-1500 words. Longer will not be graded as better: be concise, with no unnecessary words. Double spaced please.
- We will talk more about drug companies and their role in the healthcare system before then.
- This is an opinion or position piece
  1. Read about the issues and decide what your opinion is
  2. Construct a well-reasoned argument to support your opinion
  3. Back up factual statements with citations

Logistics of paper

- Include an introduction that lays out a clear thesis.
- Tie everything throughout your paper to back to this thesis
- Do not spend too much time reviewing facts that don't inform the thesis
- You may use a lead paragraph before your introduction

More Logistics

- Please use in-text citations and a bibliography (no footnote citations)
  - Use quotation marks for any passage taken word for word
- Post a word count
- Draft due April 15
  - You will receive feedback on your draft 1 week later, then have 1 week to do revisions
- Final version due May 1

Class quiz

The ISIS-2 trial had four groups of patients. Druin Burch thinks that this aspect of the experiment was ethically dubious. Why?

Any updates on class projects?

Healthcare system finances from different points of view

- Patient's view: out-of-pocket vs. insurance policy vs. employment vs. taxes
- Doctor's view
- Drug company's view

How doctors get paid is a separate question from who foots the bill

1. Doctors earn a fixed salary, e.g.
   - the Mayo clinic or Kaiser (private sector)
   - Military docs and Veterans Administration (government)
2. Fee for service
   - Most insurance (private sector)
   - Medicare and Medicaid (government)

- Fee for service gives incentives to provide more of the most profitable services, less of everything else

Healthcare system finances from different points of view

- Patient's view: out-of-pocket vs. insurance policy vs. employment vs. taxes
- Doctor's view: salary vs. fee-for-service
- Drug companies' view

Drugs are not a free market

- In exchange for doing R&D, government grants 2 kinds of monopolies
  - Patent on a drug (20 years, INCLUDING time taken for clinical trials)
  - monopoly on marketing rights (FDA)
- Monopolies mean that no competitor can sell the same thing (eg more cheaply)
- Extending monopoly for a few extra years can make a huge difference to profits, dwarfing the extensive legal costs

How are new drugs approved?

- New Drug Application (NDA) to FDA
  - includes test data on both animals and humans
  - how the drug was manufactured
- Is the drug is safe and effective? Do its benefits outweigh its ~~risks~~ harms?
- Is the drug's labeling information appropriate?
- Are the manufacturing methods adequate for ensuring the purity and integrity of the drug?

Where do new drugs come from?

Research pipeline

1. Basic research on disease mechanisms
2. Preclinical development
   - Identifying drug candidates
   - Studying them in animals and cell cultures
3. Clinical testing

Pre-clinical development

- Drug is synthesized and purified
- Animal tests
- Institutional review boards assess the studies and make recommendations on how to proceed.
- If the recommendations are positive, apply to the FDA in order to begin clinical trials.

Phase I/II trials

- Dose-finding (how much before toxicity?)
- Pharmacokinetics
  - relationship between time and plasma concentration of a drug
  - what the body does to the drug
- Pharmacodynamics
  - relationship between drug concentration and effect
  - what the drug does to the body/target

Clinical trials

- Phase 1: The drug is tested in a few healthy volunteers
- Phase 2: Various doses of the drug are tried to determine how much to give to patients
- Phase 3: Double-blind placebo controlled trials to demonstrate that the drug works
  - Comparison to best existing treatment is not always required
  - Can be MUCH larger than Phases I and II
- Phase 4: Post-approval trials that are sometimes a condition attached by the FDA to the approval
  - Sometimes a way to market to doctors and get them used to prescribing a particular drug

Where do the clinical trial patients come from?

- Access to patients must come through doctors
- NIH sponsors some trials in academic medical centers
- Drug companies once did the same, now mostly use for-profit contract research organizations
  - Establish networks of doctors who they pay to administer drugs and collect data
  - Advertise directly for patients
- From 1994 to 2004, CRO-run research increased from $1.6 to $7.6 billion while the share of clinical trials run at academic medical centers dropped from 63% to 26%

Shortage of human subjects

This, not FDA roadblocks, is the main delay in getting new drugs to market

- most patients with the condition in question don't meet trial eligibility criteria
- many trials competing to recruit the same few eligible patients

Most clinical trials are funded by industry, conducted through CROs

- Money corrupts: one survey found that industry-sponsored research was 4 times more likely to be favorable to a drug than NIH-funded research

Bias creeps in through study design

- placebo or best alternative?
  - is the alternative given at appropriate dose?
- Changing and cherry-picking outcome measures
  - Is the trial long enough?
  - Is the trial stopped early as soon as the favored drug is winning?
- Mining for subsets of patients
  - If I told you aspirin doesn't prevent heart attacks in Geminis and Libras, would you believe me?
  - What if I said it works in men not women?
- Using young subjects (who have fewer side effects) even for drugs to treat the old

Class quiz

What is the drug rep's strategy for dealing with "acquiescent" doctors, who nod and agree and seem to go along, but don't change their later prescribing behavior in the way the sales rep wants them to?

Pharma monitoring doctors: update from 2007 reading

In 2011, the Supreme Court struck down Vermont's ban on pharmacies selling prescription information for drug marketing purposes, on 1st amendment grounds

- New Hampshire and Maine had similar measures

Monitoring pharma-doctor relationships

- Physician Payments Sunshine Act (part of Affordable Care Act) requires all drug and medical-device firms have to publicly report all payments to physicians, including food, beginning Aug 1, 2013
- Official website will provide public access to the data starting Sep 30, 2014.
- Meantime, to see if your doctor received drug company money, check www.propublica.org/series/dollars-for-docs

Any updates on class projects?

Healthcare system finances from different points of view

- Patient's view: out-of-pocket vs. insurance policy vs. employment vs. taxes
- Doctor's view
  - Read current news stories: following a legal battle by the owner of the Wall Street Journal, Medicare released data yesterday on doctor reimbursements
- Drug company's view

Ways that study designs chosen by drug companies can create bias

- placebo or best alternative?
  - is the alternative given at appropriate dose?
- Changing and cherry-picking outcome measures
  - Is the trial long enough?
  - Is the trial stopped early as soon as the favored drug is winning?
- Mining for subsets of patients
  - If I told you aspirin doesn't prevent heart attacks in Geminis and Libras, would you believe me?
  - What if I said it works in men not women?
- Using young subjects (who have fewer side effects) even for drugs to treat the old

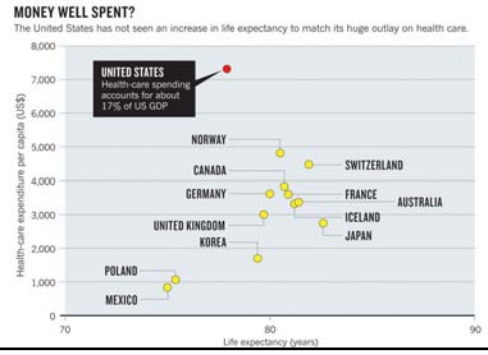
Unless we have ALL the data, there is bias

- Drug companies are allowed to do 10 trials and only publish the 2 ones that get the desired results
- They must show all data to FDA, which requires the drug to be better than placebo in 2 clinical trials
- The SSRI =80% placebo data was analyzed only after a Freedom of Information request to the FDA
- Independent researchers have been fighting Roche for years to get the data on Tamiflu

Follow the money

- Increasing prescription drug costs are a major contributor to rising healthcare costs, which, through Medicare, are a large contributor to the longterm government budget deficit
- Drugs are much more expensive in the US than in other countries
  - Medicare prescription benefit 2003 law bans government from bargaining for lower prices
  - Medicare Prescription Drug Price Negotiation Act passed the House in 2007, but has never become law. Reintroduced in 2010, 2011, 2013
  - Also illegal to buy from Canada or Mexico, although not usually prosecuted

Sobering view of the need for health care reform



2001 numbers on 10 US drug companies

- Profits = 18.5% of sales
- Research & development ~ 14% (but some of this is marketing in disguise)
  - In 2011, Pfizer slashed R&D, saying that it made more sense, and was less risky, to buy in drugs developed by academic scientists
- "Marketing and administration" = 35%
  - Conventional advertising
  - Medical "education"
  - Legal costs
  - Executive salaries
  - Political lobbying

Direct to consumer advertising

- TV ads expensive and often misleading: FDA only had 30 reviewers to cull 24,000 ads in 2001.
- May be replaced with something just as bad: Pfizer received 4 letters in 4 years about misleading Lipitor ads.
- Direct to consumer advertising is banned in all developed countries except the US and New Zealand

Bribes to doctors: 2001

- $11 billion "free samples" to get you hooked
- 88,000 gift-bearing sales reps to doctors' offices:
  - one per 5 or 6 physicians
  - one per 2.5 targeted physicians

Serious, outright bribery

- Lupron is a hormone treatment for prostate cancer
- Injected in doctor's offices, 80% paid by Medicare
- Doctors buy it directly, then bill Medicare for reimbursement based on average wholesale price
- In 1990s, competition started with cheaper Zoladex
- TAP Pharmaceuticals inflated listed price to $500, then sold it to doctors for $350: doctors kept the difference
- Company used taxpayer money to bribe doctors to prescribe their more expensive drug
- One HMO was offered a $25K "educational" grant to use however they liked: instead alerted authorities and taped conversations, settled for $875 million
- Rival later paid $355 million to settle similar charges

Just-released Medicare data hints at similar story for macular degeneration

- Lucentis costs $2,000 per monthly injection
- Cancer drug Avastin (derived from same antibody, also made by Genentech) is an alternative at about $50 for an equal dose.
- Avastin is off-label for this purpose, but there are trials showing Avastin works just as well
- Doctors get a percentage, and rumors of a rebate
- Lucentis costs Medicare about $1 billion / year

Medical "education"

- "Informing" doctors about alternative uses for a drug is legally different from "marketing" it for that use
- Laws against kickbacks have exemption for educational or research activities
- Doctors are required to receive continuing medical education (CME)
- In 2009, drug companies paid over half CME costs

Medical conferences

- Run by for-profit medical education and communication companies (MECCs)
  - some are even owned by large ad agencies
- It has been shown that doctors prescribe more of the sponsor's drugs after their meetings
- Doctors go on expensive junket to listen to talks. In exchange for minimal feedback, companies pay them as consultants.
- Speakers are paid even more. Drug companies will even make their slides for them…
- Over 300,000 pseudo-educational events in 2000, about a quarter of which offered CME credits

Thought leaders

- = prominent experts, often med school faculty, who write papers, contribute to textbooks, give talks
- Drug companies flatter them that their expertise is invaluable, and pay them handsomely as consultants.
- Eg, head of Brown University Psychiatry made over $500,000 in one year consulting for drug companies that make antidepressants

NEJM Editorial "Is academic medicine for sale?"

"No. The current owner is very happy with it."

Letter to the editor

Drug company advertising strategies

- Direct to consumer
- Sales reps to doctors
- Kickbacks to doctors
- "Education" for doctors
- "Research" for doctors

Do poor quality phase IV trials to "educate" doctors to prescribe off-label

- Absurdly little required of (paid) doctors, other than to prescribe a particular drug
- Poor quality, never submitted to FDA, not published or only in marginal journals
- In contrast, companies often drag feet over FDA-required Phase IV trials to monitor drugs for long-term side effects: they have nothing to gain here

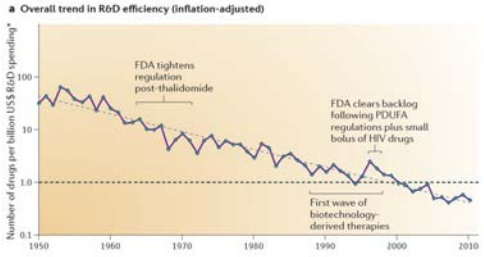
Do poor quality phase IV trials to "educate" doctors to prescribe off-label

Neurontin was approved by the FDA in 1994 for epilepsy, to be used after other drugs fail in a patient

- Patent was expiring in 1998 (extended to 2000)
- "Publications strategy" to sponsor minimal research, write favorable journal articles, pay academic "authors" to put their names on them
- Promote "findings" through "education"
- Sales of $2.7 billion in 2003
- In 2004, Pfizer pled guilty to illegal marketing and settled for $430 million
  - some of which funded today's reading
  - whistleblower received nearly $27 million

Class quiz

Name 3 of the proposed possible primary causes of the decline in pharmaceutical industry R&D efficiency. Briefly explain the meaning of ONE of them.



Proposed causes

1. Better than the Beatles
2. (Low hanging fruit)
3. Cautious regulator
4. Throw money at it
5. Basic research-brute force bias

Industry is in trouble

- Few new innovative drugs in the pipeline
- Many blockbusters are coming off patent

Trace Gleevec's history (poster child for rational and innovative drug design)

- Miracle drug for chronic myelogenous leukemia (now other cancers too)
- U Penn researcher found weird chromosome
- Other academic work found a cancer enzyme
- Chemists in Israel and Novartis synthesized molecules to inhibit enzyme
- Novartis patented several in 1994
- Brian Druker at Oregon Health & Sciences University tested them, one had very specific action in cell culture
- Novartis showed little interest, but eventually agreed to limited clinical tests by Druker
- After spectacular preliminary results in 1999, Novartis invested in large-scale trials, completed with FDA approval within 2 years
- $92,000 per year, $4.7 billion sales in 2012
- Druker plus 118 other cancer experts cosigned a complaint about the high prices of this and subsequent similar drugs

Novartis owned Gleevec patent

The Bayh-Dole Act (1980) lets universities and small businesses patent discoveries coming from NIH-funded research, then grant exclusive licenses to drug companies

Epogen history

- Hormone (erythropoietin) to treat anemia in patients with kidney failure
- Discovered 1976 at University of Chicago by Eugene Goldwasser
- NIH-funded research at Columbia University invented and patented a synthesis technique
- Small biotech startup (Amgen) licensed it to develop large-scale commercial synthesis
- U.S. 2009 cost was $8,447 per patient
- Amgen made $2 billion in 2010 selling it to Medicare
- Taxpayers paid twice, Goldwasser got nothing

Taxol history

- Blockbuster cancer drug
- All research conducted at or sponsored by National Cancer Institute over nearly 30 years, costing taxpayers $183 million
- Derived from bark of Pacific yew tree in 1960s, short supply
- In 1991, Bristol-Myers Squibb signed agreement to supply Taxol, and in 1992 was given 5 years marketing rights
- In 1994, NIH funded researchers at Florida State University devised a way to synthesize Taxol
- Licensed to Bristol-Myers Squibb in return for royalties
- $10,000-$20,000 for 1 year treatment (~20 times cost)
- Sales peaked in 2000 at $1.6 billion (generic now available)
- tens of millions to Florida State

AZT was the first drug for HIV/AIDS

- Synthesized in 1964 by Michigan Cancer Foundation
- No good for cancer, but German lab found it effective against mouse viruses in 1974
- Burroughs Wellcome acquired it for possible use against herpes
- AIDS identified 1981, HIV in 1983
- NIH screened antiviral agents, found AZT worked in test tubes and early clinical trials
- Burroughs Wellcome patented it for AIDS and did the later trials for rapid FDA approval in 1987

Burroughs Wellcome

"did not develop or provide the first application of the technology for determining whether a drug like AZT can suppress live AIDS virus in human cells, nor did it develop the technology to determine at what concentration such an effect might be achieved in humans. Moreover, it was not first to administer AZT to a human being with AIDS, nor did it perform the first clinical pharmacology studies in patients. It also did not perform the immunological and virological studies necessary to infer that the drug might work… All of these were accomplished by the staff of the National Cancer Institute working with the staff of Duke University…Indeed one of the key obstacles to the development of AZT was that Burroughs Wellcome did not work with live AIDS virus nor wish to receive samples from AIDS patients."

Sam Broder, NIH, replying to a self-congratulatory letter to the New York Times by Burroughs Wellcome CEO

Biotech industry (and reductionist approaches) did better in Avastin / Lucentis history

- In 1971, Judah Folkman at Harvard Medical School proposed fighting cancer by blocking a "tumor angiogenesis factor"
- VEGF isolated in 1989 by 2 groups, including one at Genentech, who allowed Napoleone Ferrara time to pursue this research direction as a side project
- Avastin is a humanized monoclonal antibody to VEGF, FDA approval 2004 for first cancer use following trials sponsored by Genentech
- Lucentis is a fragment of the antibody, designed to be smaller with shorter half-life, predicted more suitable for eye

How good are Avastin / Lucentis?

- NHS in UK restricts refuses to buy Avastin for most cancers, because benefit is only a few months' survival vs. high costs
- The FDA has rescinded approval for breast cancer and the NIH is currently sponsoring more than 30 Avastin trials to learn more
- While cancer results are not amazing, both drugs are pretty miraculous for wet age-related macular degeneration. Independent researchers found that both drugs work equally well.

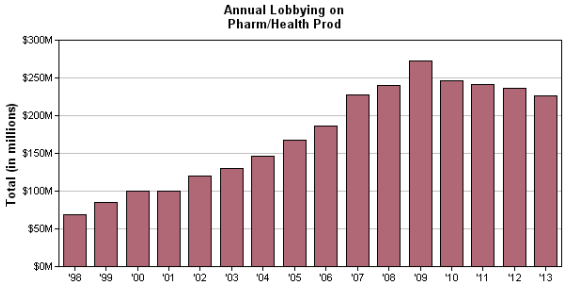
Role of pharmaceutical industry

The good
- Taking promising drugs, shepherding them through the regulatory system, organizing their manufacture and distribution
- Important work, but not the major innovation
- Occasionally hitting a drug target like VEGF esp. for biologics rather than small drugs

The bad
- Advertising to doctors and patients to promote overuse
- Promoting off-label drugs that range from useless to harmful
- Doing bad science whenever they can get away with it
- Exploiting monopolies to add pointless costs to healthcare, including lots of legal actions

The ugly
- Political influence

One of the largest lobbies in Washington



See http://www.opensecrets.org/industries/indus.php?cycle=2014&ind=H04 for more information

Questionable laws

- 1987: bans anyone but the manufacturer from importing prescription drugs from another country: even if the drugs were made in the US
- 1997: authorized an industry-supported private company (Drugdex) to decide whether Medicaid would pay for off-label drugs. It approves twice as many uses as the 2 other federally recognized (non-profit) directories
- Foot-dragging over generic HIV drugs in poor countries

What pharmaceutical companies do

"If I'm a manufacturer and I can change one molecule and get another 20 years of patent rights, and convince physicians to prescribe and consumers to demand the next form of Prilosec, or weekly Prozac instead of daily Prozac, just as my patent expires, then why would I be spending money on a lot less certain endeavor, which is looking for brand-new drugs?"

Dr Sharon Levine
Associate Executive Medical Director
Kaiser Permanente Medical Group

Prilosec / Nexium

- Prilosec is a blockbuster heartburn drug that went off-patent in 2001
- Mixture of active and perhaps inactive isomer of omeprazole
- AstraZeneca patented the active form as Nexium
- Promoted it (heavy advertising, initially cheaper) as an improvement to Prilosec in time to switch patients just before Prilosec patent expired
- Nexium sales over $5.6 billion in 2012

Most FDA approvals are "me-too drugs"

- Ie, no better than drugs already on the market to treat the same condition
- FDA only requires that drug be shown to be better than placebo, unless placebo is much more dangerous than existing treatment
- Even if a difference were found, doses may not be equivalent: maybe the right conclusion is a change in approved dose
- FDA grants exclusive marketing rights for a new use of the drug, *even if* the use is obvious

Statins lower cholesterol

- We have 7: atorvastatin (Lipitor), fluvastatin (Lescol), lovastatin (Mevacor, Altocor), pitavastatin (Livalo), pravastatin (Pravachol), rosuvastatin (Crestor) and simvastatin (Zocor)
- Sometimes tested for slightly different outcomes in different types of patients, then promoted as especially effective for those uses: other statins were not tested for that use

Selective serotonin reuptake inhibitors

- Prozac was first, developed mainly on the basis of research done outside Eli Lilly
- When it went off-patent in 2001, more expensive me-too Paxil and Zoloft used more
- Eli Lilly renamed Prozac "Sarafem", colored it pink and lavender, got FDA approval to market it for "premenstrual dsyphoric disorder"

Vioxx

- Treatment for arthritic pain
- Approved under priority review in 1998
- No study comparing it to aspirin or similar
- Theory predicted Vioxx would be easier on the stomach, although problems with data supporting this prediction
- Heavily promoted drug, 2 million took it, yearly sales of $2.5 billion

Phase IV trial

- Trial to see if Vioxx could prevent the recurrence of colorectal polyps
- Ostensibly serendipitous finding that Vioxx doubled the risk of heart attacks and strokes
- Merck then withdrew Vioxx in 2004
- In fact, there had been earlier signs of trouble, eg a 2000 trial showing Vioxx had equal pain relief as OTC naproxen (Aleve), but half the rate of serious gastrointestinal problems
- Also 4 times the heart attacks: not in the published paper but shown by later FDA analysis

What about the me-toos?

- Same cox-2 inhibitor drug class included Pfizer's Celebrex and Bextra, with 2 more me-too drugs in development
- Celebrex was first and even bigger blockbuster
- Within months, reports that Celebrex and Bextra posed similar risks
- Pfizer did not withdraw, only stopped direct to consumer advertising

What did the FDA do?

- Appointed special advisory panel, public hearings, emotional testimonials
- Recommended all 3 be on market with strong warnings
- Many panel members had financial ties to Merck or Pfizer
- FDA announced only Celebrex (which may have lower effective dose) could stay on market
- Same result as if votes from those with financial ties had been excluded

Celebrex today

- 2013 sales of $1.93 billion in US, $2.92 billion world-wide
- No better at pain relief than common drugs like ibuprobren
- Still serious questions about whether it is easier on the stomach
- Trial on heart risks began in 2006 but is scheduled to end in September 2015
- Celebrex comes off patent May 2014 (after losing a battle to extend the patent by 18 months)

Class quiz

Instead of agreeing to take part in a clinical trial, a patient can often get the same treatment outside the trial. Given the high promise of the new drug in preclinical testing and phase I and II trials, it is obviously tempting to be guaranteed the new drug instead of risking a placebo. What do the statistics say in the case of childhood cancer? On average, are patients better or worse off if they agree to be in a trial and risk getting a placebo?

Any updates on class projects?

Is breast cancer something that younger women should worry about?

Which of the two following questions should you ask?

1. What is the probability of dying of breast cancer, given that you are a young woman?
2. What is the probability of being a young woman, given that you are dying of breast cancer?

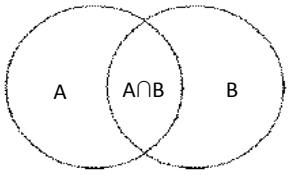
Conditional Probability

Probability of an event ***given*** some other information

Notation

P(A|B) means probability of A given B

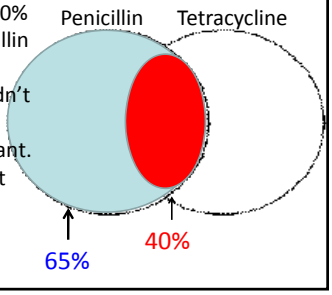
$$P(A|B)=\frac{P(A\cap B)}{P(B)}$$



Venn diagrams help visualize conditional probability

Bacteria causing disease X are regularly tested for resistance to the antibiotics penicillin and tetracycline. 65% are resistant to penicillin and 40% are resistant to both penicillin and tetracycline. You have disease X, and penicillin didn't work on you because your infection is penicillin resistant. What is the probability that your infection is also tetracycline resistant?

=0.4/0.65=0.615



That was Prob(tet|pen).
What about Prob(pen|tet)?

Assuming 100% are resistant to at least one antibiotic

Blue area alone?

25%

White area alone?

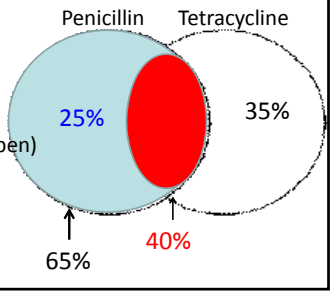
35%

Prob(pen|tet)?

40/(40+35)=53%

Not the same as prob(tet|pen)

40/(40+25)=61.5%



Real data on breast cancer frequency

1.What is the probability of being 30-49, given that you are a woman dying of breast cancer?

2.What is the probability that breast cancer is killing you, given that you are a 30-49 year old woman and dying?

3.What is the probability of dying of breast cancer in the next 20 years, given that you are a 30 year old woman?

Age	Alive at beginning of interval	Incidents of breast cancer	Deaths from breast cancer	Deaths from cardiovascular causes	Deaths from other causes
0–9	1,000	0	0	0	7
10–19	993	0	0	0	2
20–29	991	0	0	0	3
30–34	988	1	0	0	2
35–39	986	3	0	0	3
40–44	983	5	1	1	4
45–49	977	8	2	1	6
50–54	968	11	3	2	11
55–59	952	12	3	5	15
60–64	929	12	3	9	25
65–69	892	14	4	16	36
70–74	836	13	5	28	51
75–79	752	11	6	52	70
80–84	624	9	6	89	95
≥85	434	5	7	224	203

1.What is the probability of being 30-49, given that you are a woman dying of breast cancer?

40 women dying of breast cancer

3 of them are 30-49

3/40 = 0.075 = 7.5%

Age	Alive at beginning of interval	Incidents of breast cancer	Deaths from breast cancer	Deaths from cardiovascular causes	Deaths from other causes
0–9	1,000	0	0	0	7
10–19	993	0	0	0	2
20–29	991	0	0	0	3
30–34	988	1	0	0	2
35–39	986	3	0	0	3
40–44	983	5	1	1	4
45–49	977	8	2	1	6
50–54	968	11	3	2	11
55–59	952	12	3	5	15
60–64	929	12	3	9	25
65–69	892	14	4	16	36
70–74	836	13	5	28	51
75–79	752	11	6	52	70
80–84	624	9	6	89	95
≥85	434	5	7	224	203

2. What is the probability that breast cancer is killing you, given that you are a 30-49 year old woman and dying?

20 30-49 year olds dying

3 of them from breast cancer

3/20 = 0.15 = 15%

Age	Alive at beginning of interval	Incidents of breast cancer	Deaths from breast cancer	Deaths from cardiovascular causes	Deaths from other causes
0–9	1,000	0	0	0	7
10–19	993	0	0	0	2
20–29	991	0	0	0	3
30–34	988	1	0	0	2
35–39	986	3	0	0	3
40–44	983	5	1	1	4
45–49	977	8	2	1	6
50–54	968	11	3	2	11
55–59	952	12	3	5	15
60–64	929	12	3	9	25
65–69	892	14	4	16	36
70–74	836	13	5	28	51
75–79	752	11	6	52	70
80–84	624	9	6	89	95
≥85	434	5	7	224	203

Prob (die of breast cancer | die young)

≠

Prob(die young|die of breast cancer)

3. What is the probability of dying of breast cancer in the next 20 years, given that you are a 30 year old woman?

988 women alive at age 30

3 of them die from breast cancer during the next 20 years

3/988 = 0.003 = 0.3%

Age	Alive at beginning of interval	Incidents of breast cancer	Deaths from breast cancer	Deaths from cardiovascular causes	Deaths from other causes
0–9	1,000	0	0	0	7
10–19	993	0	0	0	2
20–29	991	0	0	0	3
30–34	988	1	0	0	2
35–39	986	3	0	0	3
40–44	983	5	1	1	4
45–49	977	8	2	1	6
50–54	968	11	3	2	11
55–59	952	12	3	5	15
60–64	929	12	3	9	25
65–69	892	14	4	16	36
70–74	836	13	5	28	51
75–79	752	11	6	52	70
80–84	624	9	6	89	95
≥85	434	5	7	224	203

Homework

- Probability reflection #5 on conditional probability (just taught) is available and due TUESDAY
- Homework #6 on Bayes' Theorem (about to be taught) is available and due THURSDAY

Probability(disease|symptoms)

You are working in an outpatient clinic where the records shows that during the past year, 10% of the walk-in patients have had sepsis. A patient walks in with a high fever, chills, and skin lesions. According to the records

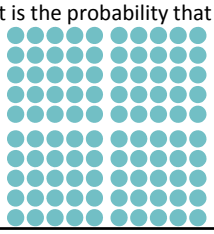
- If a patient has sepsis, there is an 80% chance that s/he will have these symptoms
- If a patient does not have sepsis, there is still a 10% chance that s/he will have these symptoms

What is the probability that this patient has sepsis?

Consider 100 patients

- 10% of walk-in patients have sepsis.
- If a patient has sepsis, there is an 80% chance that s/he will have these symptoms
- If a patient does not have sepsis, there is still a 10% chance that s/he will have these symptoms

What is the probability that this patient has sepsis?



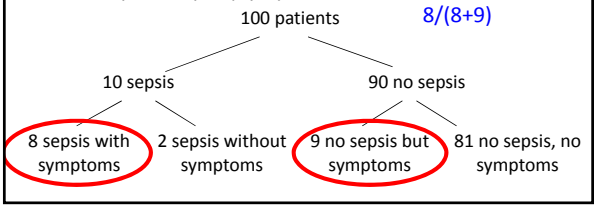
Out of 100, there are

- 8 symptomatic patients with sepsis
- 9 symptomatic patients without
- Prob(sepsis|symptoms)= 8/(8+9)

Another way of solving it

- 10% of walk-in patients have sepsis.
- If a patient has sepsis, there is an 80% chance that s/he will have these symptoms
- If a patient does not have sepsis, there is still a 10% chance that s/he will have these symptoms

What is the prob(sepsis|symptoms)?



Bayes' Theorem

- Forward probability (causal version): omniscient writer of exam questions tells you how the world is and asks you to calculate the probability of something happens
- Real world (science): You observe some data (eg symptoms), and want to figure out how it happened (eg diagnosis) i.e. probability that the world is a certain way. In this case you can use Bayes' Theorem.

Bayes' Theorem reduces this to forward probabilities

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$$

B = observed data

A and A-complement = alternative explanations for it

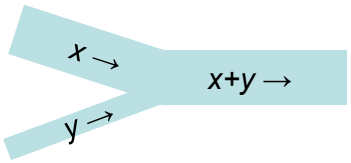
That equation makes no sense to me: I don't know how to use it and I am going to forget it

- Imagine the probability of a certain outcome happening a certain way is *x*
- The probability of the same outcome happening a different way is *y*
- These are the only two ways to get this outcome
- Now I tell you the outcome really happened, and ask you what is the probability that A rather than B is true
- $x/(x+y)$

$$P(A|B)=\frac{P(B|A)P(A)}{P(B|A)P(A)+P(B|A^c)P(A^c)}$$

Another way to look at it

- Imagine 2 or more tributaries join up to make a river
- Take one molecule of water from the river
- How likely is it to have come from each of the tributaries?
- It depends on how big each tributary is, relative to the total



Which question requires Bayes' Theorem?

Before going on vacation for a week, you ask your spacey friend to water your ailing plant. Without water, the plant has a 90 percent chance of dying. Even with proper watering, it has a 20 percent chance of dying. And the probability that your friend will forget to water it is 30 percent.

1. What's the chance that your plant will survive the week? *Forward probability*
2. If it's dead when you return, what's the chance that your friend forgot to water it? *Backward probability (Bayes')*
3. If your friend forgot to water it, what's the chance it'll be dead when you return? *Easy: 90% was given in the question*

Class quiz

What was the role of doctors in George Washington's final illness and death?

Other updates on class projects?

Which question requires Bayes' Theorem?

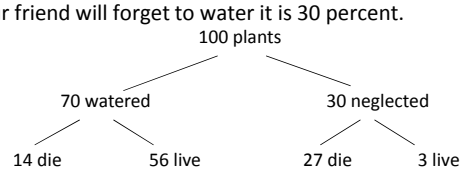
Before going on vacation for a week, you ask your spacey friend to water your ailing plant. Without water, the plant has a 90 percent chance of dying. Even with proper watering, it has a 20 percent chance of dying. And the probability that your friend will forget to water it is 30 percent.

1. What's the probability that your plant will survive the week? *Forward probability*
2. If it's dead when you return, what's the probability that your friend forgot to water it? *Backward probability (Bayes')*
3. If your friend forgot to water it, what's the probability it'll be dead when you return?

Easy forward probability: 90% was given in the question

Tree method: start with 100 plants

Before going on vacation for a week, you ask your spacey friend to water your ailing plant. Without water, the plant has a 90 percent chance of dying. Even with proper watering, it has a 20 percent chance of dying. And the probability that your friend will forget to water it is 30 percent.



1. What's the chance that your plant will survive the week?
   (56+3)/100=59%
2. If it's dead when you return, what's the chance that your friend forgot to water it? 27/(27+14)=66%

What does "probability" mean when there is only one plant? It was either watered or it wasn't!

Earlier this semester we compared

- Classic definition
  - eg each number is equally likely when you roll an ideal "fair" dice
- Frequency definition
  - you need to estimate it from data as the number of things that happened divided by the number of times it could have happened it
- Bayesian definition
  - how confident you are
  - if there *were* 100 dead-plant situations and you accused your friend of neglect, you would be right 66 times

Your turn

About 0.01% of men with no known risk factors have HIV. HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. A man with no known risk factors applies for a Green Card, and is made to take a compulsory HIV test. He tests positive. What is the probability that he has HIV?

Hint: instead of using 0.01%, 99.9% etc, consider 10,000 men, using a tree.

Percentages

- 0.01% of 10,000: you shouldn't need a calculator for this!
- Remember that "per cent" means "per hundred"

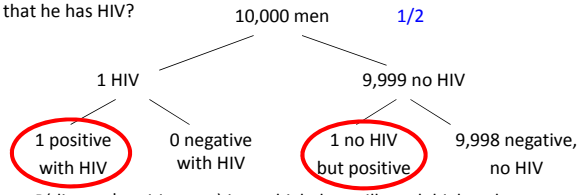
Your turn

About 0.01% of men with no known risk factors have HIV. HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. A man with no known risk factors applies for a Green Card, and is made to take a compulsory HIV test. He tests positive. What is the probability that he has HIV?

Hint: instead of using 0.01%, 99.9% etc, consider 10,000 men, using a tree.

Tree method

About 0.01% of men with no known risk factors have HIV. HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. A man with no known risk factors tests positive. What is the probability that he has HIV?



- P(disease|positive test) is not high, but still much higher than original base rate P(disease).
- Note that P(disease) goes up even though test doesn't CAUSE disease.

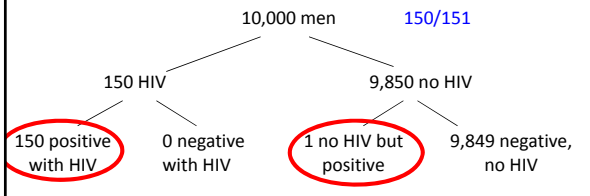
The base rate is important

Now consider a homosexual man. The rate of HIV in this population is 1.5%.

As before, HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. What is the probability that this man has HIV, given that he tests positive?

Higher base rate

The rate of HIV in this population is 1.5%. HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. What is the probability that this man has HIV, given a positive test?

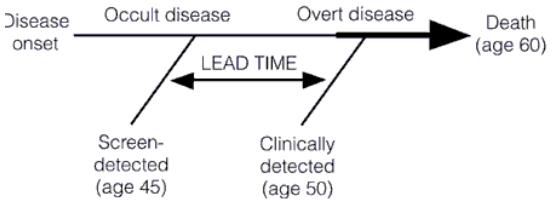


Mammograms to detect breast cancer

- All women over a certain age are encouraged to have regular mammograms
- The idea is to detect breast cancer earlier, before a woman notices a lump
- It is believed that early detection leads to more effective treatment, before the tumor progresses

Lead time bias

- What if early detection doesn't help?
- How long after her diagnosis-day would the woman below survive with vs. without early detection?



Use Mortality Rate instead of Survival Time

"Let's say there's a new cancer of the thumb killing people. From the time the first cancer cell appears, you have nine years to live, with chemo. From the time you can feel a lump, you have four years to live, with chemo. Let's say we have no way to detect the disease until you feel a lump. The five year survival rate for this cancer is about 0, because within five years of detection, everyone dies, even on therapy.

Use Mortality Rate instead of Survival Time

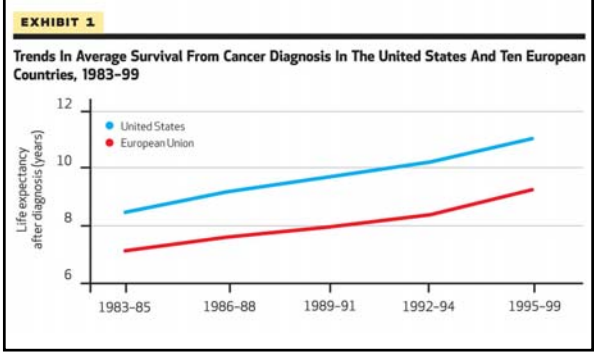
Now I invent a new scanner that can detect thumb cancer when only one cell is there. Because it's the United States, we invest heavily in those scanners. Early detection is everything, right? We have protests and lawsuits and now everyone is getting scanned like crazy. Not only that, but people are getting chemo earlier and earlier for the cancer. Sure, the side effects are terrible, but we want to live.

Use Mortality Rate instead of Survival Time

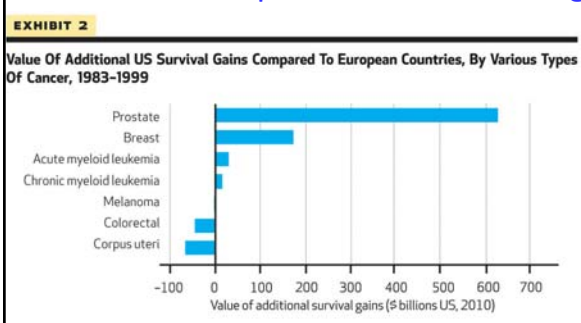
We made no improvements to the treatment. Everyone is still dying four years after they feel the lump. But since we are making the diagnosis five years earlier, our five year survival rate is now approaching 100%! Everyone is living nine years with the disease. Meanwhile, in England, they say that the scanner doesn't extend life and won't pay for it. Rationing! That's why their five year survival rate is still 0%."

- Aaron Carroll

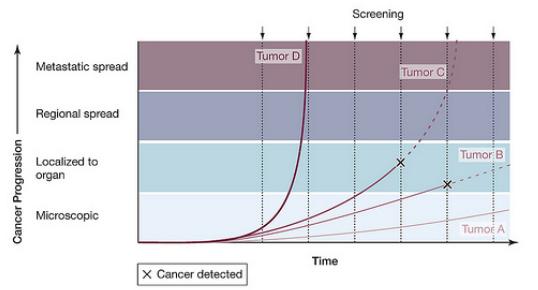
In Europe, most women get mammograms every 2 years between the ages of 50 and 70. In the US, screening often begins at 40, is annual, and may not stop.



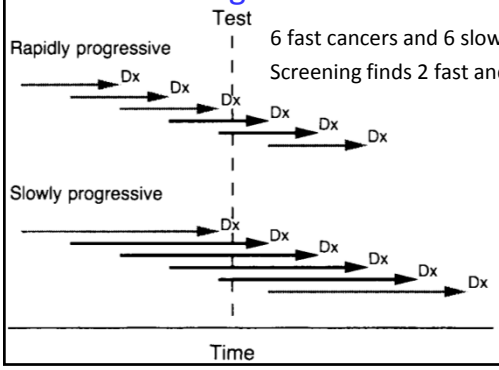
Survival Gains come from cancers associated with questionable screening



Length bias: some cancers grow faster than others



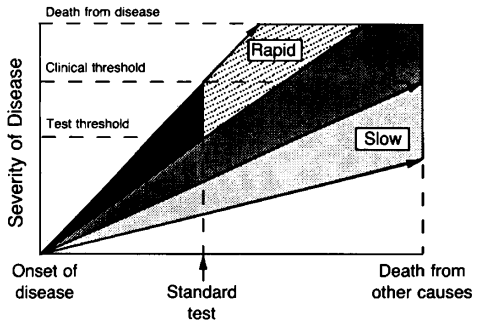
Length bias: some cancers grow faster than others



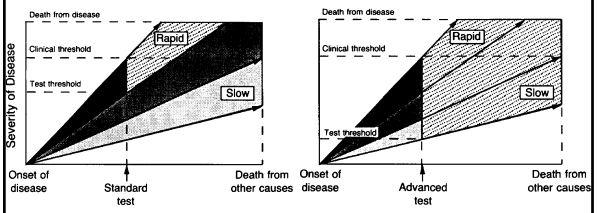
6 fast cancers and 6 slow.

Screening finds 2 fast and 4 slow.

What happens if you invent a "better", more sensitive test?



More sensitive early-detection tests catch more slow-growing cancers, making length bias worse



Mammograms have harms

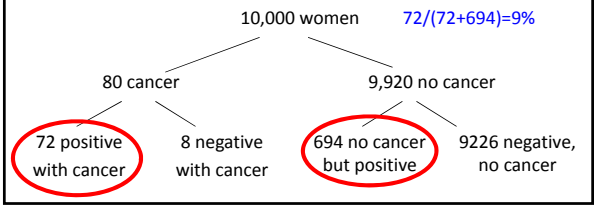
- After 10 mammograms, there is a 61% risk of a false positive, leading to emotional distress and perhaps a biopsy before learning that it is a false positive

Mammograms to detect breast cancer

0.8% of women aged between 40 and 50 have breast cancer. If a woman has breast cancer, there is a 90% chance of seeing it in a mammogram. If she does NOT have breast cancer, there is still a 7% chance of a false positive mammogram. Imagine a woman who has a positive mammogram, after routine screening. What is the probability that she has breast cancer?

Tree method

0.8% of women have breast cancer. If a woman has breast cancer, there is a 90% chance of seeing it in a mammogram. If she does NOT have breast cancer, there is still a 7% chance of a false positive mammogram. Imagine a woman who has a positive mammogram. What is the probability that she has breast cancer?



Class quiz

What is the most important fact about a drug that cannot be found in advertisements, or sometimes even in the FDA-approved information insert?

Updates on class projects?

Mammograms have harms

- Radiation
- After 10 mammograms, there is a 61% risk of a false positive, leading to emotional distress and perhaps a biopsy before learning that it is a false positive

Overdiagnosis

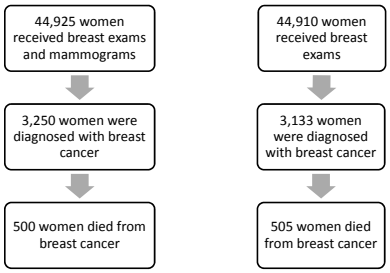
- Some cancers may not progress at all
- Overdiagnosis means finding something that would never have bothered the patient: pseudodisease
- Among 40-50 year old women, 40% will have microscopic evidence of breast cancer if you look hard enough (based on autopsies): danger as new imaging technologies become more sensitive

Should asymptomatic women get mammograms?

- Lead time bias, length bias, and overdiagnosis all make early detection look good based on survival times
- Randomized trials are the only way to really know
- Most were done a long time ago, and breast cancer treatments have improved a lot since then

- Some BRCA mutations have lifetime cancer risk as high as 80%: these women are different. So are women who detect a lump.

Most recent trial, done in Canada



- No significant benefit
- One overdiagnosed woman per 424 screened

Meta-analysis of many studies (some old) suggests some benefit, but also much harm

Table 2. Estimated Benefits and Harms of Mammography Screening for 10 000 Women Who Undergo Annual Screening Mammography Over a 10-Year Period

Age, y	No. Diagnosed With Invasive Breast Cancer or DCIS During the 10 y of Screening[a]	No. of Breast Cancer Deaths in next 15 y[b]	No. of Deaths Averted With Mammography Screening Over Next 15 y[c]	No. of Breast Cancers or DCIS Diagnosed During the 10 y That Would Never Become Clinically Important (Overdiagnosis)[d]	No. (95% CI) With ≥1 False-Positive Result During the 10 y[e]	No. (95% CI) With ≥1 Unnecessary Biopsy During the 10 y[e]
40	190	27-32	1-16	?-104[f]	6130 (5940-6310)	700 (610-780)
50	302	56-64	3-32	30-137	6130 (5800-6470)	940 (740-1150)
60	438	87-97	5-49	64-194	4970 (4780-5150)	980 (840-1130)

- Key benefit?
- Key harm?
- Harms vs. benefits?
  - 40-50: ~2 to 104 overdiagnoses per 1 life saved
  - 50-60: ~1 to 46 overdiagnoses per 1 life saved
  - 60-70: ~1 to 39 overdiagnoses per 1 life saved
- For what harm : benefit ratio would you get screened?

"I'm so grateful I had a mammogram and my cancer was caught early"

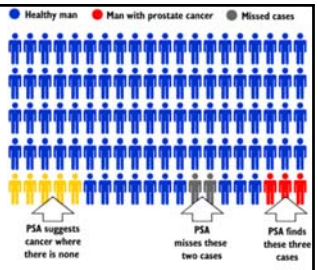
- Prob(overdiagnosis|screening diagnosis) ~ 0.19
  - i.e., 19% of such "grateful patients" are the people *most harmed* by mammography
- Even if it wasn't an overdiagnosis, in some cases prognosis would be no different if diagnosed later

Prostate cancer

- 70% of dead men over 80 have prostate cancer, as seen by autopsy
  - it never bothered them enough to be diagnosed, and they died of something else
  - diagnosis would have harmed them
- When we screen, we can't tell which cancers are the bad ones
- No evidence that PSA screening reduces all-cause mortality

Prostate cancer

- Nice figure (found online): what is wrong with it?
- How often does early diagnosis change the prognosis of the 3 men? What are the side effects of their treatment?



Pap smears

- I'm not aware of randomized trials
- After screening became routine in the US, rates of cervical cancer decreased from 14.2 per 100 000 (1973) to 7.8 per 100 000 (1994)
- Most deaths are in women who have not been screened
- Recent changes to guidelines suggest that certain classes of women should be screened less often: nearly the same benefit, much less harm

Class quiz

Name two medical treatments that are so obviously beneficial that they require no evidence from randomized trials

Updates on class projects?

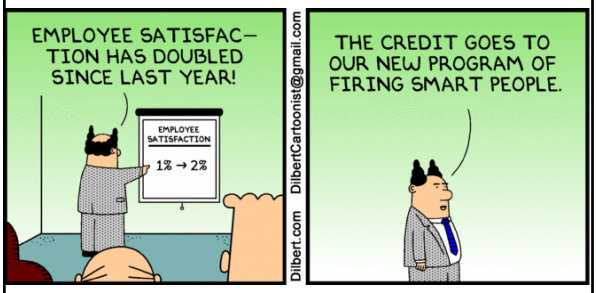
Relative vs. absolute scales

Which is a better way to measure the impact on your budget?

1. Gas prices doubled (or went up by 20% etc.)
2. Gas prices rose by $1 per gallon (or 50c per gallon etc.)

The same relative vs. absolute distinction applies to frequencies and probabilities

Relative vs absolute frequencies



Clinical significance depends on absolute risk

Do not confuse

- statistical significance ($p < 0.05$)
- clinical significance (absolute effect size is big enough to care)

Different ways to communicate benefits and/or harms. Best-estimate: mammography ages 50-60 lowers death rates from ~7 per 1000 to ~6 per 1000

- Relative risk reduction: reduces risk of dying by one seventh (14%)
- Absolute risk reduction 1 out of 1,000 (0.1%)
- Number needed to treat (NNT) in order save one life = 1000
- Increase in life expectancy: average of ~12 days
  = driving ~300 fewer miles / year

How do you communicate if you are

1. Selling mammograms?
2. Deciding what to do as a doctor?
3. As a patient?

Best-estimate number needed to harm (NNH) by overdiagnosis: 175

Problem set #7 due May 6

- Includes not just the last topic of relative vs. absolute risk, NNT etc but ALSO revision of other topics taught throughout the semester
- Probability reflection Q7 due before next class May 1
- Bonus probability reflection Q8 encouraged for May 6. If you have missed a due date for a past probability reflection, here is your chance to make it up!
- Bonus homework set #8 (on material I am about to teach) can substitute for one you did badly on, email to Parris >24 hours before final exam

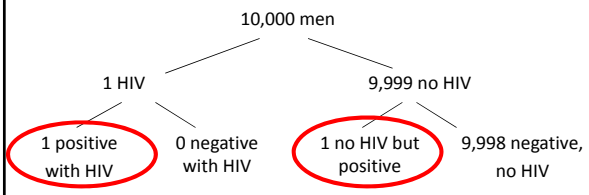
How do we decide between

1. Null hypothesis: treatment is useless
2. Alternative hypothesis: treatment works

- Is the *p*-value < cutoff α (normally 0.05)?
- What is a *p*-value?
  Probability(concluding something | nothing there)
- Is this the right question to ask?
  - What we really want is
    Prob(something there|*p*-value summarizing our data)
  - Positive predictive value (PPV)
- Does this sound familiar?

Bayes theorem gives us a PPV (1/2 in this eg)

About 0.01% of men with no known risk factors have HIV. HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. A man with no known risk factors tests positive. What is the probability that he has HIV?

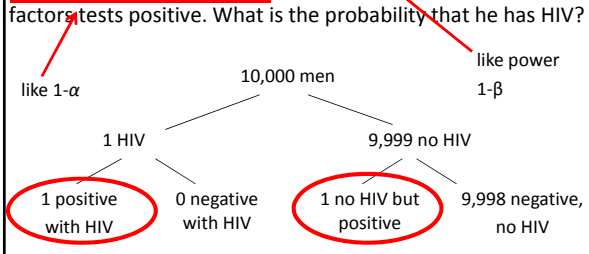


Two different kinds of error

		Null hypothesis	
		True (no diff)	False (difference)
Decision	Accept null (no diff) p > 0.05	Correct	Type II Error β False negative
	Reject null (diff) p < 0.05	Type I Error α False positive	Correct (Power=1-β)

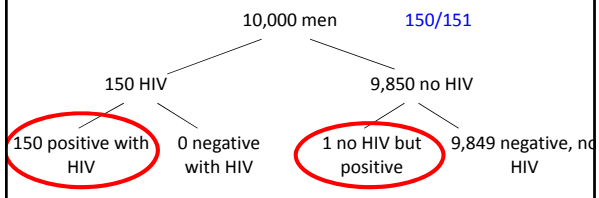
What are α and β?

About 0.01% of men with no known risk factors have HIV. HIV+ men test positive 99.9% of the time. HIV- men test negative 99.99% of the time. A man with no known risk factors tests positive. What is the probability that he has HIV?



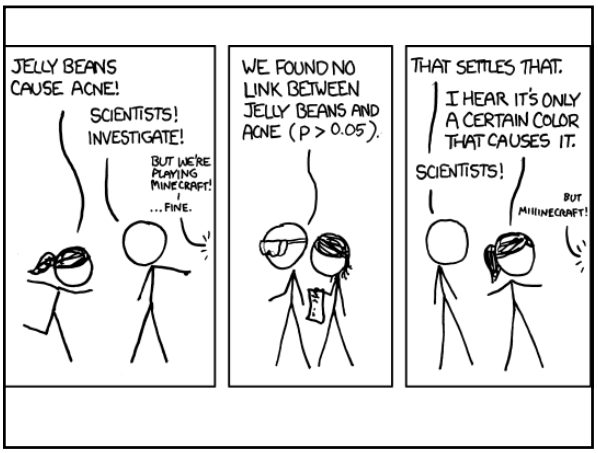
The base rate / prior probability affects the PPV but not α, β or *p*

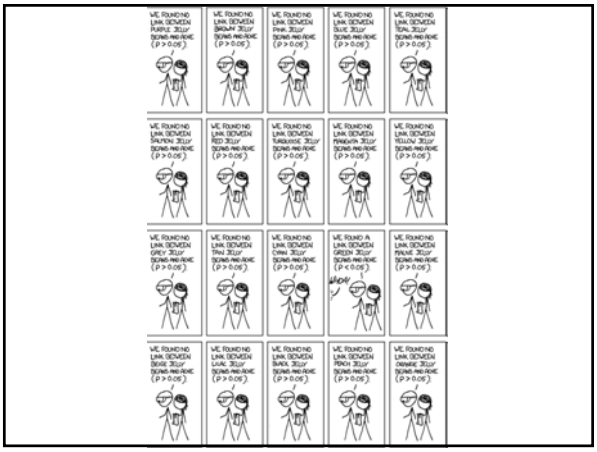
Now consider a homosexual man. The rate of HIV in this population is 1.5%.

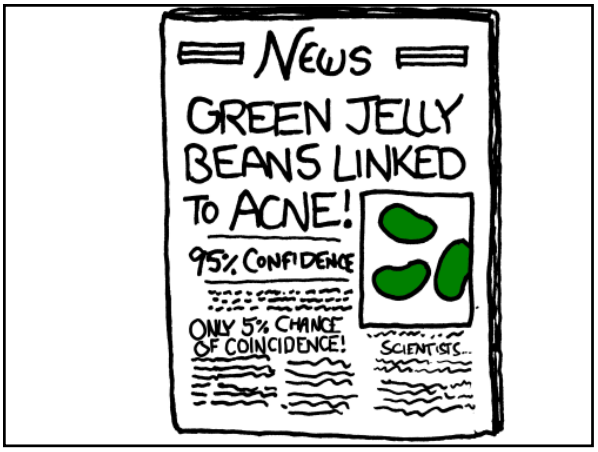


Finding a *p*-value < 0.05 is like a positive test on a screen

- Base rate: how many of the hypotheses tested by scientists are true?
- Let's take an extreme example: all the hypotheses are false.
- Then all the ones with $p < 0.05$ are false too.





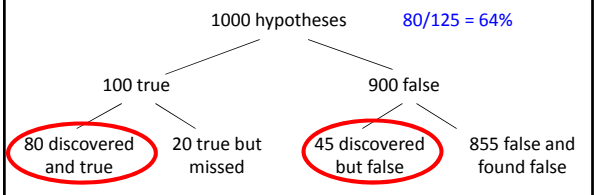


What proportion of published scientific results are true?

- Let's say 1 in 10 hypotheses is true
- *p*-value < cutoff $\alpha = 0.05$
- Power = $1-\beta = 80\%$ likely to find evidence for a true hypothesis

What proportion of published scientific results are true?

- Let's say 1 in 10 hypotheses is true
- *p*-value < cutoff $\alpha = 0.05$
- Power = $1-\beta = 80\%$ likely to find evidence for a true hypothesis



Why Most Published Research Findings Are False

John P. A. Ioannidis

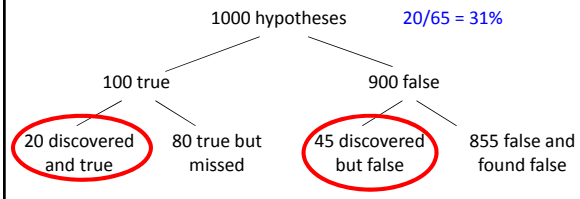
Summary

There is increasing concern that most current published research findings are false. The probability that a research claim is true may depend on study power and bias, the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field. In this framework, a research finding is less likely to be true when the studies conducted in a field are smaller; when effect sizes are smaller; when there is a greater number and lesser preselection of tested relationships; where there is greater flexibility in designs, definitions, outcomes, and analytical modes; when there is greater financial and other interest and prejudice; and when more teams are involved in a scientific field in chase of statistical significance. Simulations show that for most study designs and settings, it is more likely for a research claim to be false than true. Moreover, for many current scientific fields, claimed research findings may often be simply accurate measures of the prevailing bias. In this essay, I discuss the implications of these problems for the conduct and interpretation of research.

It can be proven that most claimed research findings are false.

What about in small, underpowered studies?

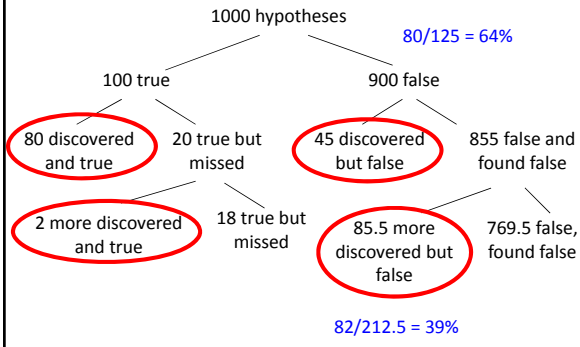
- Still assume 1 in 10 hypotheses is true
- *p*-value < cutoff $\alpha = 0.05$
- Now power = 1-β = 20% likely to find evidence for a true hypothesis



What if the research is biased?

- Still assume 1 in 10 hypotheses is true
- *p*-value < cutoff $\alpha = 0.05$
- Power = 1-β = 80% likely to find evidence for a true hypothesis
- BUT 1/10 studies that should have been negative were badly conducted and/or analyzed so that they came out positive

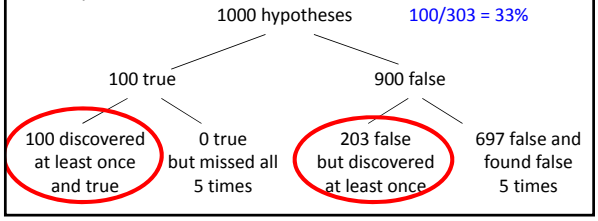
Add 10% bias. How many published results are true?



82/212.5 = 39%

What if 5 research groups test the same hypothesis, and only publish positive results?

- Probability that false hypothesis will test false 5 times is
- i.e. α is effectively 0.226 not 0.05
- Probability that true hypothesis will test false 5 times is β^5 =



When are research results most likely to be true?

- When they were likely to be true, even before the research was done (prior plausibility)
- When only one possible relationship is tested, rather than reporting the best of many
- When the study has high power
  - Lots of subjects
  - Large effect size
- When there is less bias, eg
  - few financial and other conflicts of interest
  - less flexibility in the data analysis
- When a similar study is unlikely to have already been tried multiple times, ie when the field is *less hot*

Last class quiz

Does a *p*-value of 0.05 mean that the null hypothesis has only a 5% chance of being true? If not, what *does* a *p*-value tell you?

Post-course assessment

- These post-tests assess the course; they will NOT be used to assess you
- Please complete the ATS now
  (and the TCEs once I leave the room)
- Write your student number on your ATS, or if you can't remember it, write your name
- Please complete the QRQ post-test on D2L
  - Your grade will not be released until this is done
  - You will get many emails from me: do it today and avoid them all!

Updates on class projects?

- Wildcat article today!
- Project leaders, when everything is done except the final report writing, please email me and Parris a list of who did what (including both members of your group, and other members of the class who volunteered as partcipants). Author contributions should also be clearly marked on the final report.

When are research results most likely to be true?

- When they were likely to be true, even before the research was done (prior plausibility)
- When only one possible relationship is tested, rather than reporting the best of many
- When the study has high power
  - Lots of subjects
  - Large effect size
- When there is less bias, eg
  - few financial and other conflicts of interest
  - less flexibility in the data analysis
- When a similar study is unlikely to have already been tried multiple times, ie when the field is *less hot*

Table 4. PPV of Research Findings for Various Combinations of Power (1 – β), Ratio of True to Not-True Relationships *(R)*, and Bias *(u)*

1 – β (power)	Base rate / prior prob. (R)	u (bias)	Practical Example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	0.85
0.95	2:1	0.30	Confirmatory meta-analysis of good-quality RCTs	0.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	0.41
0.20	1:5	0.20	Underpowered, but well-performed phase I/II RCT	0.23
0.20	1:5	0.80	Underpowered, poorly performed phase I/II RCT	0.17
0.80	1:10	0.30	Adequately powered exploratory epidemiological study	0.20
0.20	1:10	0.30	Underpowered exploratory epidemiological study	0.12
0.20	1:1,000	0.80	Discovery-oriented exploratory research with massive testing	0.0010
0.20	1:1,000	0.20	As in previous example, but with more limited bias (more standardized)	0.0015

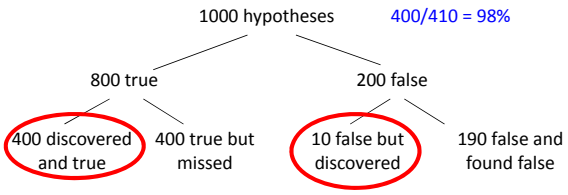
The estimated PPVs (positive predictive values) are derived assuming $\alpha = 0.05$ for a single study.
RCT, randomized controlled trial.
DOI: 10.1371/journal.pmed.0020124.t004

p-value cutoff (type I error rate)

What would you think of the following evidence and why?

1. a large, well-conducted, placebo-controlled trial of homeopathy shows effectiveness against migraines ($p < 0.05$)
2. after anecdotes that a new painkiller drug reduced anxiety, the patent owner confirmed this with a large placebo-controlled trial ($p < 0.05$)
3. a study looking at associations between schizophrenia and different genes across the entire genome, finding 20 genes that have $p < 0.001$
4. a small, placebo-controlled study ($p < 0.05$) showing the effectiveness of a weekly slow-release version of an antihistamine previously known to be effective against allergies when taken daily

80% likely to be true? 50% power?



How am I supposed to come up with a "prior probability"?

- Bayesian definition of probability
  - how confident you are
  - if there *were* 100 dead-plant situations and you accused your friend of neglect, you would be right 66 times BUT there is only one plant
  - Bayesian prior probability (base rate) meets evidence and then turns into a posterior probability (PPV)
- Classic definition
  - eg each number is equally likely when you roll an ideal "fair" dice
- Frequency definition
  - you need to estimate it from data as the number of things that happened divided by the number of times it could have happened it

We can use Bayes theorem to *calculate* how we expect many published results to be false, but how can we *test* this?

How do published results stand the test of time?

Contradicted and Initially Stronger Effects in Highly Cited Clinical Research

John P. A. Ioannidis, MD

Context Controversy and uncertainty ensue when the results of clinical research on the effectiveness of interventions are subsequently contradicted. Controversies are most prominent when high-impact research is involved.

Objectives To understand how frequently highly cited studies are contradicted or find effects that are stronger than in other similar studies and to discern whether specific characteristics are associated with such refutation over time.

Design All original clinical research studies published in 3 major general clinical journals or high-impact-factor specialty journals in 1990-2003 and cited more than 1000 times in the literature were examined.

Main Outcome Measure The results of highly cited articles were compared against subsequent studies of comparable or larger sample size and similar or better controlled designs. The same analysis was also performed comparatively for matched studies that were not so highly cited.

Results Of 49 highly cited original clinical research studies, 45 claimed that the intervention was effective. Of these, 7 (16%) were contradicted by subsequent studies, 7 others (16%) had found effects that were stronger than those of subsequent studies, 20 (44%) were replicated, and 11 (24%) remained largely unchallenged. Five of 6 highly-cited nonrandomized studies had been contradicted or had found stronger effects vs 9 of 39 randomized controlled trials ($P=.008$). Among randomized trials, studies with contradicted or stronger effects were smaller ($P=.009$) than replicated or unchallenged studies although there was no statistically significant difference in their early or overall citation impact. Matched control studies did not have a significantly different share of refuted results than highly cited studies, but they included more studies with "negative" results.

Conclusions Contradiction and initially stronger effects are not unusual in highly cited research of clinical interventions and their outcomes. The extent to which high citations may provoke contradictions and vice versa needs more study. Controversies are most common with highly cited nonrandomized studies, but even the most highly cited randomized trials may be challenged and refuted over time, especially small ones.

JAMA. 2005;294:218-228 www.jama.com

CLINICAL RESEARCH ON IMPORTANT questions about the efficacy of medical interventions is sometimes followed by subsequent studies that either reach opposite conclusions or suggest that the original claims were too strong. Such disagreements may upset clinical practice and acquire publicity in both scientific circles and in the lay press. Several empirical investigations have tried to address whether specific types of studies are more likely to be contradicted and to explain observed controversies. For example, evidence exists that small studies may sometimes be refuted by larger ones.[1,2]

Similarly, there is some evidence on disagreements between epidemiological studies and randomized trials.[3-7] Prior investigations have focused on a variety of studies without any particular attention to their relative importance and scientific impact. Yet, most research publications have little impact while a small minority receives

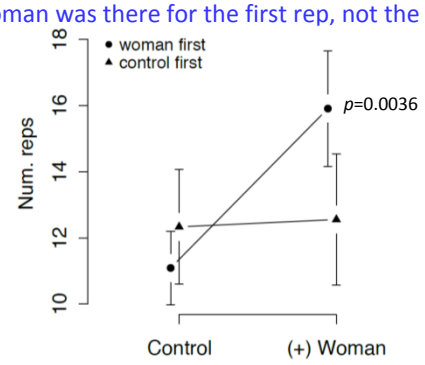
My office hours reminder: tomorrow 11:30-12:30 LSS327a

The final exam will not repeat previous questions, but will test your ability to transfer what you have learned to new applications

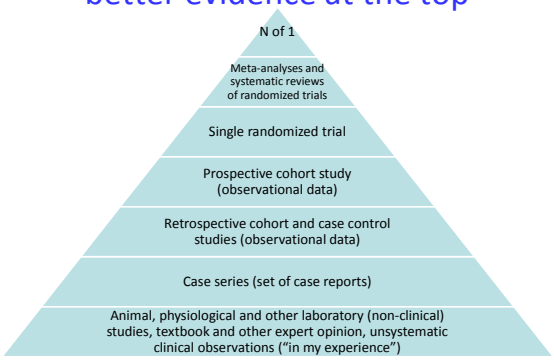
Please complete the QRQ post-test on D2L

- Thanks to the 14/40 completed so far!
- These post-tests assess the course; they will NOT be used to assess you

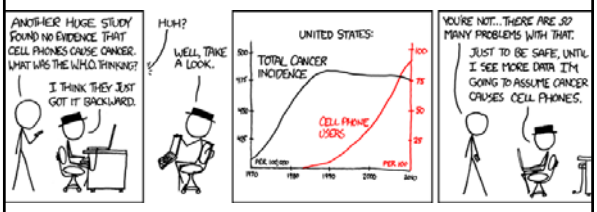
Weightlifting project:
Having the woman there helped BUT ONLY IF the woman was there for the first rep, not the second



Evidence Pyramid: better evidence at the top



Observational studies measure correlation. You need an intervention and a control group to measure causation



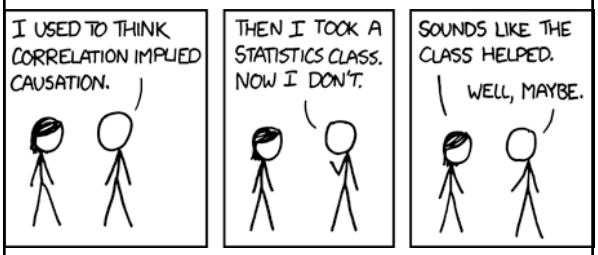
I encourage you all to find examples of this in news stories

Do high-voltage power lines damage your health?

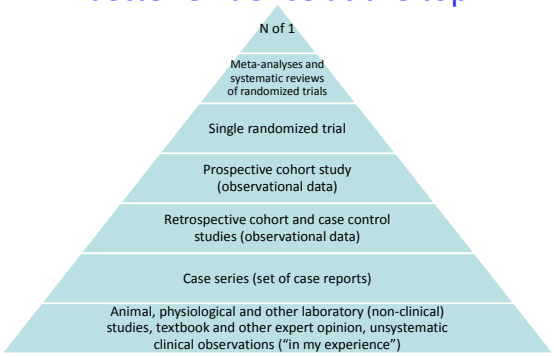
Correlation with unhealthiness

- BUT these homes are unpopular and thus cheaper
- poorer people live in cheaper homes
- poorer people have worse health

Observational studies measure correlation.
You need an intervention and a control group to measure causation



Evidence Pyramid: better evidence at the top



POSTMENOPAUSAL ESTROGEN AND PROGESTIN USE AND THE RISK OF CARDIOVASCULAR DISEASE

Background Estrogen therapy in postmenopausal women has been associated with a decreased risk of heart disease. There is little information, however, about the effect of combined estrogen and progestin therapy on the risk of cardiovascular disease.

Methods We examined the relation between cardiovascular disease and postmenopausal hormone therapy during up to 16 years of follow-up in 59,337 women from the Nurses' Health Study, who were 30 to 55 years of age at base line. Information on hormone use was ascertained with biennial questionnaires. From 1976 to 1992, we documented 770 cases of myocardial infarction or death from coronary disease in this group and 572 strokes. Proportional-hazards models were used to calculate relative risks and 95 percent confidence intervals, adjusted for confounding variables.

Results We observed a marked decrease in the risk of major coronary heart disease among women who took estrogen with progestin, as compared with the risk among women who did not use hormones (multivariate adjusted relative risk, 0.39; 95 percent confidence interval, 0.19 to 0.78) or estrogen alone (relative risk, 0.60; 95 percent confidence interval, 0.43 to 0.83). However, there was no significant association between stroke and use of combined hormones (multivariate adjusted relative risk, 1.09; 95 percent confidence interval, 0.66 to 1.80) or estrogen alone (relative risk, 1.27; 95 percent confidence interval, 0.95 to 1.69).

Conclusions The addition of progestin does not appear to attenuate the cardioprotective effects of postmenopausal estrogen therapy. (N Engl J Med 1996;335:453-61.)

Effects of Conjugated Equine Estrogen in Postmenopausal Women With Hysterectomy

The Women's Health Initiative Randomized Controlled Trial

Context Despite decades of use and considerable research, the role of estrogen alone in preventing chronic diseases in postmenopausal women remains uncertain.

Objective To assess the effects on major disease incidence rates of the most commonly used postmenopausal hormone therapy in the United States.

Design, Setting, and Participants A randomized, double-blind, placebo-controlled disease prevention trial (the estrogen-alone component of the Women's Health Initiative [WHI]) conducted in 40 US clinical centers beginning in 1993. Enrolled were 10739 postmenopausal women, aged 50-79 years, with prior hysterectomy, including 23% of minority race/ethnicity.

Intervention Women were randomly assigned to receive either 0.625 mg/d of conjugated equine estrogen (CEE) or placebo.

Main Outcome Measures The primary outcome was coronary heart disease (CHD) incidence (nonfatal myocardial infarction or CHD death). Invasive breast cancer incidence was the primary safety outcome. A global index of risks and benefits, including these primary outcomes plus stroke, pulmonary embolism (PE), colorectal cancer, hip fracture, and deaths from other causes, was used for summarizing overall effects.

Effects of Conjugated Equine Estrogen in Postmenopausal Women With Hysterectomy

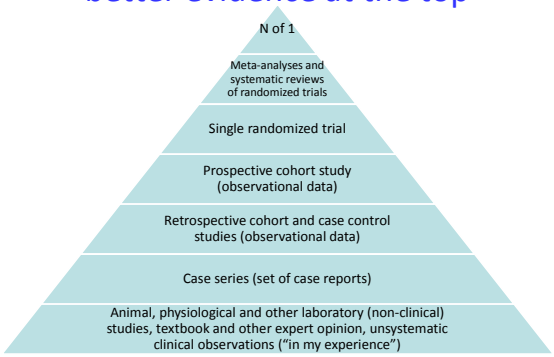
The Women's Health Initiative Randomized Controlled Trial

Results In February 2004, after reviewing data through November 30, 2003, the National Institutes of Health (NIH) decided to end the intervention phase of the trial early. Estimated hazard ratios (HRs) (95% confidence intervals [CIs]) for CEE vs placebo for the major clinical outcomes available through February 29, 2004 (average follow-up 6.8 years), were: CHD, 0.91 (0.75-1.12) with 376 cases; breast cancer, 0.77 (0.59-1.01) with 218 cases; stroke, 1.39 (1.10-1.77) with 276 cases; PE, 1.34 (0.87-2.06) with 85 cases; colorectal cancer, 1.08 (0.75-1.55) with 119 cases; and hip fracture, 0.61 (0.41-0.91) with 102 cases. Corresponding results for composite outcomes were: total cardiovascular disease, 1.12 (1.01-1.24); total cancer, 0.93 (0.81-1.07); total fractures, 0.70 (0.63-0.79); total mortality, 1.04 (0.88-1.22), and the global index, 1.01 (0.91-1.12). For the outcomes significantly affected by CEE, there was an absolute excess risk of 12 additional strokes per 10000 person-years and an absolute risk reduction of 6 fewer hip fractures per 10000 person-years. The estimated excess risk for all monitored events in the global index was a nonsignificant 2 events per 10000 person-years.

Conclusions The use of CEE increases the risk of stroke, decreases the risk of hip fracture, and does not affect CHD incidence in postmenopausal women with prior hysterectomy over an average of 6.8 years. A possible reduction in breast cancer risk requires further investigation. The burden of incident disease events was equivalent in the CEE and placebo groups, indicating no overall benefit. Thus, CEE should not be recommended for chronic disease prevention in postmenopausal women.

JAMA, 2004;291:1701-1712 www.jama.com

Evidence Pyramid: better evidence at the top



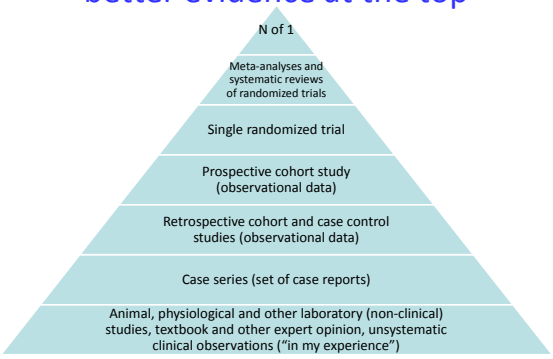
Effect of fat saturation on satiety, hormone release, and food intake

Background: Ileal delivery of fat reduces hunger and food intake through activation of the ileal brake. Physicochemical properties of fat have been shown to affect satiety and food intake.
Objective: The objective of this study was to assess the effect of ileal fat emulsions with differing degrees of fatty acid saturation on satiety, food intake, and gut peptides (cholecystokinin and peptide YY). We hypothesized that long-chain triacylglycerols with diunsaturated fatty acids would increase satiety and reduce energy intake compared with long-chain triacylglycerols with monounsaturated or saturated fatty acids.
Design: We performed a double-blind, randomized, crossover study in which 15 healthy subjects [mean age: 24 y; mean body mass index (in kg/m^2): 22] were intubated with a naso-ileal catheter and participated in 4 experiments performed in random order on 4 consecutive days. After consumption of a liquid meal, subjects received a fat or control infusion in the ileum. Fat emulsions consisted of 6 g of 18:0 (shea oil; mainly 18:0), 18:1 (canola oil; mainly 18:1), or 18:2 (safflower oil; mainly 18:2) oils. Food intake was measured during an ad libitum lunch. Satiety questionnaires (visual analog scale) and blood samples were collected at regular intervals.

Effect of fat saturation on satiety, hormone release, and food intake

Results: Compared with the control, only 18:2 and 18:1 significantly increased fullness and reduced hunger. No effect on food intake was observed. 18:1 and 18:2 increased cholecystokinin secretion significantly compared with the control. Fatty acid saturation did not affect peptide YY secretion.
Conclusions: When infused into the ileum, triacylglycerols with unsaturated fatty acids increase satiety, whereas triacylglycerols with saturated fatty acids does not. This trial was registered with the Dutch Trial Register as ISRCTN51742545. *Am J Clin Nutr* 2009;89:1019–24.

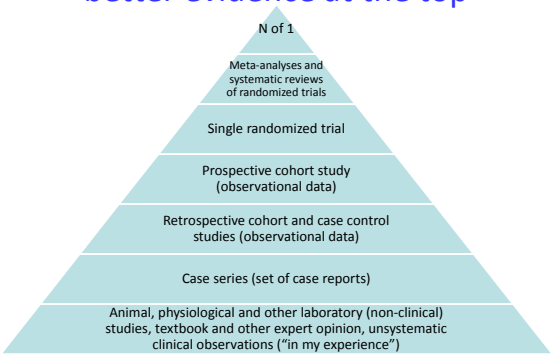
Evidence Pyramid: better evidence at the top



Meta-analysis of prospective cohort studies evaluating the association of saturated fat with cardiovascular disease[1–5]

Background: A reduction in dietary saturated fat has generally been thought to improve cardiovascular health.
Objective: The objective of this meta-analysis was to summarize the evidence related to the association of dietary saturated fat with risk of coronary heart disease (CHD), stroke, and cardiovascular disease (CVD; CHD inclusive of stroke) in prospective epidemiologic studies.
Design: Twenty-one studies identified by searching MEDLINE and EMBASE databases and secondary referencing qualified for inclusion in this study. A random-effects model was used to derive composite relative risk estimates for CHD, stroke, and CVD.
Results: During 5–23 y of follow-up of 347,747 subjects, 11,006 developed CHD or stroke. Intake of saturated fat was not associated with an increased risk of CHD, stroke, or CVD. The pooled relative risk estimates that compared extreme quantiles of saturated fat intake were 1.07 (95% CI: 0.96, 1.19; $P = 0.22$) for CHD, 0.81 (95% CI: 0.62, 1.05; $P = 0.11$) for stroke, and 1.00 (95% CI: 0.89, 1.11; $P = 0.95$) for CVD. Consideration of age, sex, and study quality did not change the results.
Conclusions: A meta-analysis of prospective epidemiologic studies showed that there is no significant evidence for concluding that dietary saturated fat is associated with an increased risk of CHD or CVD. More data are needed to elucidate whether CVD risks are likely to be influenced by the specific nutrients used to replace saturated fat. *Am J Clin Nutr* 2010;91:535–46.

Evidence Pyramid: better evidence at the top



[Intervention Review]

Reduced or modified dietary fat for preventing cardiovascular disease

Background

Reduction or modification of dietary fat can improve total cholesterol levels, but may also have a variety of effects, both positive and negative, on other cardiovascular risk factors.

Objectives

The aim of this systematic review was to assess the effect of reduction or modification of dietary fats on total and cardiovascular mortality and cardiovascular morbidity over at least 6 months, using all available randomized clinical trials.

Search strategy

The Cochrane Library, MEDLINE, EMBASE, CAB Abstracts, CVRCT registry and related Cochrane Groups' trial registers were searched through spring 1998, SIGLE to January 1999. Trials known to experts in the field and biographies were included through May 1999.

Selection criteria

Trials fulfilled the following criteria: 1) randomized with appropriate control group, 2) intention to reduce or modify fat or cholesterol intake (excluding exclusively omega-3 fat interventions), 3) not multi factorial, 4) healthy adult humans, 5) intervention at least six months, 6) mortality or cardiovascular morbidity data available. Inclusion decisions were duplicated, disagreement resolved by discussion or a third party.

Data collection and analysis

Rate data were extracted by two independent reviewers and meta-analysis performed using random effects methodology. Meta-regression and funnel plots were used.

[Intervention Review]

Reduced or modified dietary fat for preventing cardiovascular disease

Main results

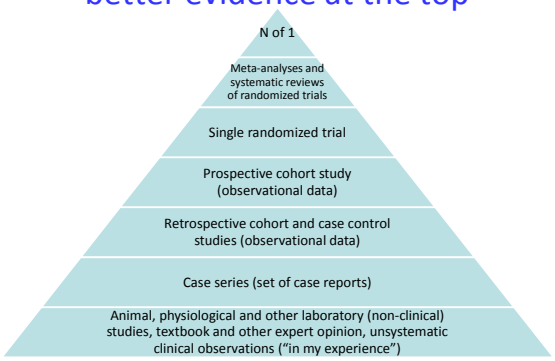
Twenty seven studies were included (40 intervention arms, 30,901 person-years). There was no significant effect on total mortality (rate ratio 0.98, 95% CI 0.86 to 1.12), a trend towards protection form cardiovascular mortality (rate ratio 0.91, 95% CI 0.77 to 1.07), and significant protection from cardiovascular events (rate ratio 0.84, 95% CI 0.72 to 0.99). The latter became non-significant on sensitivity analysis.

Trials where participants were involved for more than 2 years showed significant reductions in the rate of cardiovascular events and a suggestion of protection from total mortality. The degree of protection from cardiovascular events appeared similar in high and low risk groups, but was statistically significant only in the former.

Authors' conclusions

The findings are suggestive of a small but potentially important reduction in cardiovascular risk in trials longer than two years. Lifestyle advice to all those at high risk of cardiovascular disease (especially where statins are unavailable or rationed), and to lower risk population groups, should continue to include permanent reduction of dietary saturated fat and partial replacement by unsaturates.

Evidence Pyramid: better evidence at the top



Evidence-Based Medicine Search Engine

http://www.ahsl.arizona.edu/ebmsearch/index.cfm