1

Introduction

1.1 Why is statistical analysis so important for clinical research?

Most treatments are not sufficiently effective for you to tell whether or not they work based solely on clinical experience. You need statistical analysis!

Consider the question of whether or not to anticoagulate patients with atrial fibrillation (a condition where the heart beats irregularly) and normal heart valves. Such patients are predisposed to emboli (blood clots that travel to other parts of the body). Although anticoagulation with warfarin prevents strokes due to emboli, it can cause serious side effects (bleeding). So what do you do if you have a patient with atrial fibrillation and normal heart valves?

I remember distinctly how Dr. Kanu Chatterjee, one of the greatest cardiologists to have ever practiced medicine, answered this question in 1987. I was among the medical residents congregated around him at University of California, San Francisco Medical Center waiting for pearls of wisdom. He took a deep breath and said: “What you do is you anticoagulate all your patients with atrial fibrillation until one of them bleeds into his head. Then you don’t anticoagulate any of your patients until one of them has a stroke. Then you go back to anticoagulating all of them.”

Dr. Chatterjee was admitting with an honesty and humility often missing in clinical medicine that it was not clear whether the benefits of anticoagulation outweighed the risks. He was also capturing the tendency of physicians to base their decisions, in the absence of definitive evidence, on their most recent experience.

Fifteen years later, a pooled analysis of six randomized clinical trials demonstrated that anticoagulation with warfarin was superior to aspirin for patients with atrial fibrillation and normal heart valves (Table 1.1).[1]

Note that the risk of ischemic stroke is lower with warfarin (2.0 events per 100 patient-years) than with aspirin (4.3 events per 100 patient-years). Although the

[1] van Walraven, C., Hart, R.G., Singer, D.E., et al. Oral anticoagulants versus aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. *J. Am. Med. Assoc.* 2002; 288: 2441–8.

Table 1.1. Should you anticoagulate persons with atrial fibrillation and normal heart values?

	Events per 100 patient-years	
	Warfarin	Aspirin
Rate of ischemic stroke	2.0	4.3
Rate of major bleed	2.2	1.3

Data from van Walraven, C., et al. Oral anticoagulants versus aspirin in nonvalvular atrial fibrillation: an individual patient meta-analysis. *J. Am. Med. Assoc.* 2002; 228: 2441–8.

Statistics are needed to quantify differences that are too small to recognize through clinical experience alone.

risk of a major bleed is higher with warfarin (2.2 events per 100 patient-years) than with aspirin (1.3 events per 100 patient-years) this increase is smaller than the decrease in ischemic strokes. No cardiologist, no matter how many patients with atrial fibrillation he or she has cared for and no matter how careful he or she is at tracking the outcomes of those patients, could recognize such small but important differences through experience alone.

Even if you had the ability to detect such small differences in clinical outcomes you would still need statistics to determine whether the detected difference was greater than the difference you would expect by chance. After all, you would not expect the experience of patients receiving anticoagulation to be exactly the same as those not receiving anticoagulation. There would be some difference. The important question is whether the difference reflects a true difference between the two groups or random (chance) variation.

To understand how statistical analysis helps us evaluate the role of chance in producing differences between groups, let us consider a familiar example: the flip of a coin.

If you flip a coin that is equally weighted on both sides a hundred times (sample size, also known as *N*, of 100) it will land on heads *about* 50 times and tails *about* 50 times. I have italicized "about" because it represents chance intruding on truth. The truth is that an equally weighted coin should produce an equal number of heads and tails. But because of chance you may not get an equal number of heads and tails. Instead you may get 51 heads and 49 tails, or 49 heads and 51 tails, or 48 heads and 52 tails, etc. None of these results would make you suspicious that the coin was more heavily weighted on one side than the other.

But if the coin lands too often on a particular side, you will get suspicious as to whether the coin really is equally weighted. At a certain point, you will conclude that the difference between the results you were expecting (50–50) and the results that the coin is producing are so great that it cannot be due to chance.

Table 1.2. What result with 100 tosses would make you believe that the coin is not equally weighted on both sides?

100 tosses		
Heads, N (%)	Tails, N (%)	Probability*
50 (50)	50 (50)	1.0
49 (49)	51 (51)	0.92
48 (48)	52 (52)	0.69
45 (45)	55 (55)	0.32
40 (40)	60 (60)	0.05
35 (35)	65 (65)	0.003

* Probability of the observed data (or a more extreme result in either direction) when the expected probability for heads/tails is 0.50.

Table 1.2 quantifies what you already know intuitively. It shows the probability of obtaining a variety of results (or a more extreme result) assuming that an equally weighted coin is flipped 100 times.

By convention, a probability (P-value) of less than 0.05 is said to be *statistically significant.*

You can see that with 100 tosses even a distribution as unequal as 45% heads and 55% tails has a good chance of being due to chance alone (0.32 or about 1 in 3 trials). This probability is too high to conclude confidently that the coin is weighted more heavily on one side. However, if you have a more disproportionate distribution of 40% heads and 60% tails the probability that the result is due to chance is markedly smaller (0.05 or about 1 in 20 trials). By convention, a probability (P-value) of less than 0.05 is said to be *statistically significant*. In other words, unlikely to be due to chance. Whether you use the conventional cut-off of $P < 0.05$ or a more or less stringent one depends in part on the harm that would come from being wrong (i.e., rejecting the null hypothesis when it is correct or accepting the null hypothesis when it is wrong).

You will find that when sample sizes are large, even small differences are statistically significant. For example, the probability of obtaining a particular result (or a more extreme one) if you flip a coin 1000 times is shown in Table 1.3. Note, that with 1000 flips, having 45% land on heads results in a low probability ($P = 0.002$) that chance is the correct explanation of the results. Compare this to Table 1.2. When we had only 100 flips we could not reject the null hypothesis with a split of 45% and 55%. This should not surprise you. With more flips (a larger sample size) you have more data on which to make a determination that the coin is not acting as you would expect it to. Therefore, with larger sample sizes smaller differences from what would be expected will tip you off that the coin is not equally weighted.

Table 1.3. What result with 1000 tosses would make you believe that the coin is not weighted equally on both sides?

1000 tosses		
Heads, *N* (%)	Tails, *N* (%)	Probability*
500 (50)	500 (50)	1.0
490 (49)	510 (51)	0.52
480 (48)	520 (52)	0.22
450 (45)	550 (55)	0.002
400 (40)	600 (60)	<0.001
350 (35)	650 (65)	<0.001

* Refer to footnote of Table 1.2.

Table 1.4. What result with ten tosses would make you believe that the coin is not weighted equally on both sides?

10 tosses		
Heads, *N* (%)	Tails, *N* (%)	Probability*
5 (50)	5 (50)	1.0
4 (40)	6 (60)	0.75
2 (20)	8 (80)	0.11
1 (10)	9 (90)	0.02
0 (0)	10 (100)	0.002

* Refer to footnote of Table 1.2.

Conversely, with small samples even large differences could occur by chance alone. For example, if you toss a coin only 10 times a 20%/80% split could occur with an equally weighted coin due to chance alone with a reasonably high frequency ($P = 0.11$ or 1 in 9 times) (Table 1.4). It is only when you reach a 10%/90% split that the probability dips below the conventional threshold for rejecting the null hypothesis ($P < 0.05$).

The coin toss example illustrates that the two key elements in determining whether a result is due to chance are (1) the magnitude of the difference from what would be expected by chance; and (2) the sample size.

The more a result differs from what would be expected by chance and the larger the sample size, the more likely it is that the result cannot be explained by chance. When a result is unlikely to be due to chance you can consider alternative

The two key elements in determining whether a result is due to chance are the magnitude of the difference from what would be expected by chance and the size of the sample.

explanations. In the case of the coin toss example, if the probability of a particular result is very low, you can consider the possibility that you are dealing with an unfair coin.

A similar process occurs when considering whether two variables are associated with one another. For example, Ponsky and colleagues assessed whether health insurance status was associated with appendiceal rupture in children.[2] Appendiceal rupture occurs when an infected appendix is not removed quickly enough. Children without private health insurance may not be taken to the doctor when they have the early mild symptoms of appendicitis because they have poor access to care.

Definition

The null hypothesis is that there is no association between two variables, or no difference between two or more groups.

To assess an association between two variables, we begin by assuming that the null hypothesis is true. The null hypothesis is that there is no association between two variables, or no difference between two or more groups. In this case, the null hypothesis is that there is no association between having private health insurance and appendiceal rupture in children.

Having stated the null hypothesis we collect data to see if we can reject the null hypothesis. The ability to reject the null hypothesis when it is false is referred to as the power of a study.

Definition

Power is the ability to reject the null hypothesis when it is false.

Ponsky and colleagues used administrative data from 36 pediatric hospitals in the USA to assess the association between having private health insurance and appendiceal rupture. They found that appendiceal rupture was less likely to occur among privately insured children (32%) than children without private insurance (44%) (Table 1.5). But is it possible that the association between insurance status and appendiceal rupture is solely due to chance sampling of the underlying population? After all, this sample of 18,312 children is just one of an infinite number of samples that could be taken of children with appendicitis.

Although each such sample would likely produce a (slightly or very) different association between insurance status and appendiceal rupture, the question we need to answer is: how likely is it that we could get the data seen in Table 1.5, if there were no true association between health insurance status and appendiceal rupture?

To answer this question we perform a chi-squared analysis (Section 5.2). The small *P*-value of the chi-squared tells you that it is very unlikely that we would have gotten a sample with the data shown in Table 1.5, if there were no association between insurance status and appendiceal rupture in the population.

[2] Ponsky, T.A., Huang, Z.J., Kittle, K., et al. Hospital- and patient-level characteristics and the risk of appendiceal rupture and negative appendectomy in children. *J. Am. Med. Assoc.* 2004; 292: 1977–82.

Table 1.5. Association of insurance status with appendiceal rupture in children

	Appendiceal rupture	
Private health insurance	Yes	No
Yes	3085 (32)	6644 (68)
No	3804 (44)	4779 (56)

Chi-squared *P*-value = 0.002.
Values represented as *N* (%).
Data from Ponsky, T.A., Huang, Z.J., Kittle, K., et al. Hospital- and patient-level characteristics and the risk of appendiceal rupture and negative appendectomy in children. *J. Am. Med. Assoc.* 2004; 292: 1977–82.

Definition

Inferential statistics are used to draw conclusions about populations from samples of those populations.

Statistics (such as the chi-squared) that are used to draw conclusions about populations from samples are referred to as inferential statistics. We infer the truth about the population from the findings in the sample.

Having eliminated chance sampling from the population as the reason for this association, we can consider the alternative explanation: that there is an association between insurance status and appendiceal rupture.

A common mistake at this point in the process is to assume that if there is an association, the association is causal (i.e., not having health insurance leads to delays in appendectomy). But causality is only one alternative explanation of an association that is not due to chance. Another alternative explanation is confounding (i.e., the apparent association between two variables is actually due to a third variable or variables, Section 2.3.A). In the case of this study, there is a possibility that low income, which is associated with insurance status, may be the true cause of the higher rate of appendiceal rupture. Another alternative explanation is reverse causality (i.e., the "effect" causes the "cause", Section 2.6.A). This is an unlikely explanation in this case, since it is hard to imagine how having appendiceal rupture would lead to not having private insurance, but reverse causality may be true in other instances. Finally, bias (systematic error in measurement due to flaws in the design and/or conduct of the study, Section 2.3.B) is an issue in all studies. For example, bias could affect the results if uninsured children with appendiceal rupture were more likely to be transferred to one of the hospitals in this sample than insured children with appendiceal rupture.

The best way to eliminate these other alternative explanations is through rigorous study design. Therefore, I have placed the chapter on study design (Chapter 2), ahead of the chapters on statistical analysis. Other strategies for strengthening causal inference are discussed in Section 9.2.

Another common mistake is to assume that your results can be generalized to (can be assumed to be true for) other populations than the one that was sampled (Section 2.4). For example, Ponsky and colleagues drew their sample from the population of children having appendectomies. Whether adults without health insurance are also more likely to have appendiceal rupture than insured adults cannot be answered by their data. (But has been answered in the affirmative by other studies![3])

[3] Braveman, P., Schaaf, V.M., Egerter, S., Bennett, T., Schecter, W. Insurance-related differences in the risk of ruptured appendix. *New Engl. J. Med.* 1994; 331: 444–9.

2

Designing a study

2.1 How do I choose a research question?

The first step in designing a study is to formulate a research question. Most clinical researchers appropriately wish to study a question in their field of practice. But knowing that you want to do a research project in a field such as HIV/AIDS or cardiology or orthopedics, is quite different than having a research question. For example: What about HIV/AIDS, are you interested in studying? Methods of preventing infection? How to diagnose infection? The prevalence of infection? Survival with HIV/AIDS? The frequency of specific HIV/AIDS manifestations?

One of the best ways to identify a research question is to determine what the unknowns are in your field. What do you and the other clinicians in your field wish you knew but don't? Perhaps your clinical experience suggests to you that a particular condition is more common in one population than another, but you're not sure if your clinical experience is typical or not. Perhaps you've evaluated a patient with a particular symptom and found that the literature lacked compelling data on how to treat the patient or what test to perform next.

Research questions may be descriptive or analytic. As implied by the name, descriptive questions focus on explaining clinical phenomena such as prevalence of disease (e.g., What is the prevalence of HIV among homeless persons?), survival trends (e.g., What is the proportion of men with prostate cancer who are alive at 5 years?), health service utilization (What is the proportion of seniors receiving influenza (flu) vaccination?), and clinical test characteristics (e.g., What is the mean value of D-dimer levels among patients who have had a venous thromboembolism?).

Analytic questions are comparative: For example: Is HIV prevalence higher among homeless persons than among housed persons? Is survival among men with prostate cancer better with surgery or radiation? Are seniors with health insurance more likely to receive flu vaccine than uninsured seniors? Are persons with higher D-dimer levels more likely to have a recurrent venous thromboembolism than patients with lower levels?

In general, analytic questions are more interesting than descriptive ones because answering them may enable us to develop interventions to prevent disease or better target interventions to particular populations. However, descriptive questions often must be answered first. For example, without a thorough understanding of the baseline frequency of a condition, it may be impossible to design a study to answer an analytic question.

Whether you are answering a descriptive or analytic question, specify the population in which you will be answering the question: men, women, elders, youth, homeless persons, etc.

In choosing a research question, remember that life is short and the time it takes to complete research projects is long. (The median time between the start of enrollment of subjects and the publication of results was found to be 5.5 years for randomized controlled efficacy trials.[4]) Choose a question for which your excitement is sufficient to sustain you through tedious protocol revisions, temperamental collaborators, protective human subjects review committees, lagging enrollment, subjects who drop out of your study, missing data, writer's block, slow journal editors, jealous reviewers, and the myriad of other obstacles to performing and publishing good research.

Try to choose a research question that will have an impact on the health and well being of a population you care about. Sometimes researchers get so caught up in the academic game of grantsmanship, publication, and promotion, that they lose sight that the purpose of clinical research is to improve health by identifying risk factors of disease, improving diagnoses, finding new treatments, etc. Much well-done health care research is published that has no impact on health care.

A turning point in my research career was a study I performed on temporal trends in AIDS-related opportunistic infections.[5] At the time, clinicians noted a change in the pattern of opportunistic infections and malignancies in patients with AIDS. Specifically, with the advent of prophylaxis for *Pneumocystis carinii* pneumonia, the rate of other opportunistic infections for which we had no form of prophylaxis at that time, such as disseminated *Mycobacterium avium* complex and cytomegalovirus were increasing. I used data from a natural history cohort to determine the rate of the different manifestations by calendar year.

From an academic point of view, the study was a success. It got accepted for publication on the first submission to the leading infectious disease journal. I had

[4] Ioannidis, J.P. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *J. Am. Med. Assoc.* 1998; 279: 281–6.

[5] Katz, M.H., Hessol, N.A., Buchbinder, S.P., et al. Temporal trends of opportunistic infections and malignancies in homosexual men with AIDS. *J. Infect. Dis.* 1994; 170: 198–202.

reason to feel pleased with myself, but I wasn't. By the time the paper appeared in print, I realized it made no discernable difference in the care of persons with HIV/AIDS. All I had done was to quantitate the rate at which people were developing (then) unpreventable infections. I vowed to myself that I would focus my future research efforts on research that was more likely to have an impact.

Of course, it is sometimes difficult to fully appreciate the impact a study will have before you do it. Also, there have been instances when a study that had no immediate impact turned out to be influential in moving a field forward many years later. Nonetheless, the chance that your work will have an impact is greater if you address an important clinical question.

Another way to ensure that the results of your study will matter is to enroll a sufficient number of subjects (Chapter 7) so that a null result is meaningful. A study that detects no difference between two groups, but does not have a sufficient sample size to rule out a meaningful difference, is of no use.

In choosing a research question, consider what questions you are in a particularly good position to answer based on the prevalence of the disease in your area, your prior experience, your colleagues, and your community contacts. It is not a coincidence that most of the research on Burkett's lymphoma is performed in Africa or that most of the research on esophageal cancer is performed in Japan.

Finally, before devoting too much time to your research question, be sure it has not already been answered. This has become significantly easier in the age of computerized literature searches. Pub Med (http://www/ncbi.nlm.nih.gov/PubMed/) is a great resource. It places the holdings of the National Library of Medicine at your fingertips, free of charge.

It is also worth consulting with others in the field to see if a similar study is underway or has been presented at a conference (unfortunately not all abstracts and/or proceedings are electronically accessible).

Although, it is rare that a single study definitively answers a question, it is much less exciting to perform a study that has already been done, unless you are sure you can do it better!

In summary, before undertaking a research project, ask yourself:
Am I truly interested in knowing the results?
Will the results have an impact on clinical practice?
Will I have enough study subjects to answer the question?
Am I in a particularly good position to answer the question?
Has this question already been answered sufficiently well?

If your answers are Yes, Yes, Yes, Yes, and No, get to work choosing a study design![6]

[6] For more on choosing a research question, see Hulley, S.B., Cummings, S.R., Browner, W.S., Grady, D., Hearst, N., Newman, T.B. *Designing Clinical Research* (2nd edition). Philadelphia: Lippincott Williams & Wilkins, 2001, pp. 17–24.