Communicating Research to the General Public

The WISL Award for Communicating PhD Research to the Public launched in 2010, and since then over 100 Ph.D. degree recipients have successfully included a chapter in their Ph.D. thesis communicating their research to non-specialists. The goal is to explain the candidate's scholarly research and its significance—as well as their excitement for and journey through their area of study—to a wider audience that includes family members, friends, civic groups, newspaper reporters, program officers at appropriate funding agencies, state legislators, and members of the U.S. Congress.

WISL encourages the inclusion of such chapters in all Ph.D. theses everywhere, through the cooperation of PhD candidates, their mentors, and departments. WISL offers awards of $250 for UW-Madison Ph.D. candidates in science and engineering. Candidates from other institutions may participate, but are not eligible for the cash award. WISL strongly encourages other institutions to launch similar programs.

Wisconsin Initiative for Science Literacy

The dual mission of the Wisconsin Initiative for Science Literacy is to promote literacy in science, mathematics and technology among the general public and to attract future generations to careers in research, teaching and public service.

Contact: Prof. Bassam Z. Shakhashiri

UW-Madison Department of Chemistry

bassam@chem.wisc.edu

www.scifun.org

A Proxy-Based Approach for Unmeasured Confounding, and Other Latent Variable Problems

by

Haley Colgate Kottler

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Mathematics)

at the
University of Wisconsin-Madison
2026

Date of Final Oral Exam: 4/22/2026
The dissertation is approved by the following members of the Final Oral Committee:

Amy Cochran, Assistant Professor, Mathematics
and Population Health Sciences
Sebastien Roch, Professor, Mathematics
Jose Israel Rodriguez, Associate Professor, Mathematics
David Anderson, Professor, Mathematics

Chapter 6

A gentle introduction to causal inference with proxy variables

I am including this chapter in my dissertation to explain my research to the general public, with the support and encouragement of Bassam Shakhashiri, Cayce Osborne, and Elizabeth Reynolds from the Wisconsin Initiative for Science Literacy. What follows will correspond to Chapter 5 of my thesis. There are two primary reasons this is important to me. First, my research was conducted with financial support from the National Science Foundation and the University of Wisconsin. Publicly funded science should serve the public. There are so many important discoveries that only happened because of public support for science, but those discoveries have not necessarily been shared in comprehensible ways. This breakdown hurts both the general populace and the scientific community. People do not trust what they do not understand, and do not fight to keep funding projects that seem entirely irrelevant to their lives. This is a small step towards mending that gap.

On a more personal level, I also want to provide the start of an answer to the question every math educator hears regularly: "When am I ever going to use this?"

6.1 Introduction

Imagine a healthcare provider is working at an emergency department when an older adult in their 70s comes in saying they've been having terrible chest pain. Someone records their vitals, measurements like heart rate, blood pressure, and oxygen saturation. The provider runs some tests to try to figure out what is wrong, but in the mean time they need to make a decision. Do they hospitalize this patient or send them home? Hospitalization means they'll have extremely quick access to help if anything happens and they'll have close monitoring, but it also means they have an increased risk of infection and too much time in bed can cause muscle loss, both of which can have significant consequences for older adults. If this were one of my grandparents, I would hope their medical team uses both their best judgment and the most up to date research to make such an important decision.

Now imagine the scientist trying to figure out the best answer to the question of whether you hospitalize an older adult with chest pain who goes to the emergency department, or send them home. The gold standard in science is an experiment. If you can change one factor and control everything else, any difference in outcomes must be because of the thing you changed. If we were to set up an experiment, we would recruit older adults who go to the emergency department with chest pain, and randomly assign them to either the treatment group, hospitalization, or the control group, going home. We could then compare the average outcome for everyone who was hospitalized against the average outcome for everyone sent home, and have a better idea of the right choice to make for these patients.

However, this would be wildly unethical. Randomizing hospital admissions would put people's lives and well-being at risk unnecessarily. Healthcare providers use their training and experience to make this decision based on the specific situation of each patient. No ethics board would approve a project like this, so experimentation is out of the question. If nothing else, I certainly would not be okay with my grandparents being sent home from the hospital based on random chance. However, emergency departments collect

a lot of information. A large emergency department will have records on hundreds, if not thousands, of older adults with chest pain, what decision the provider made, and what happened with that patient. We can learn from this observational (rather than experimental) data without putting patients at risk.

There's just one catch. Typically, the patients who are hospitalized are sicker, or showing more concerning symptoms, than the patients who are sent home. If we compare the outcomes for everyone who was hospitalized against everyone who was sent home, and they were worse, we wouldn't be able to tell if it was because they were hospitalized or because they were sicker. This problem is called confounding, because it confounds the relationship between the intervention we care about (hospitalization) and it's effect on the outcome. In particular, since we can't directly measure how "sick" a patient actually is, this is unmeasured confounding. My work focuses on creating new strategies for dealing with unmeasured confounding using proxies, which are indirect measures of the confounder. For example, your blood pressure could be a proxy for how sick you are.

The idea of using proxies to handle unmeasured confounding came from a series of papers my advisor and some collaborators wrote where they modeled hospital admissions for older adults with chest pain in the emergency department[76, 77, 78, 79] (which is also why that is my go to example). The traditional methods all suggested that hospitalization makes it more likely that a patient will end up back in the emergency department. However, once they included an extra step that categorized patients as either high health needs or low health needs, to approximate the unmeasured confounder of how "sick" a patient is, they found that for low health needs patients hospital admission increases risk, while for high health needs patients hospital admission decreases risk. Essentially, they could compare the more sick hospitalized patients against the more sick discharged patients, and the less sick hospitalized patients against the less sick discharged patients, and see that the effect is different for each group. That fits better with what medical practitioners see in practice, so including that extra step made their analysis better, and could improve other projects too.

Their technique was very specific to hospital admissions, so while it worked well in that case, it was not obvious how to extend this strategy to other projects. However, a similar strategy for addressing confounding could improve decision making in a lot of situations where we cannot do experiments. Going forward, I'll explain what the usual approach to confounding has been, and when it can go wrong. Then, I'll introduce my new strategy, and show why it is an improvement on the usual approach. Finally, I'll describe the case it doesn't address yet, and what the plan is to try to fix that.

I will at some point have some numerical examples (I'm a mathematician, I can't resist it. I'm sorry!) but I promise I'll explain everything in words too, and keep the numbers as nice as possible. I will be fabricating examples to illustrate my points with nice and clean math, using a baseball analogy so that I am not misleading you with incorrect health information.

Side note: baseball is a mathematician's dream sport! Everything from pitch speeds to the number of bases stolen is measured and recorded. There's even a book about how baseball scouters combine intuition and statistics to find the right players: *Scouting and Scoring* by Christopher J. Phillips!

6.2 Background

We will start by assuming that there is no unmeasured confounding. Given the sheer amount of data we collect and have access to with electronic health records, this sometimes seems like a reasonable assumption to make. If the confounders are all measured and recorded, we can correct for their influence in a relatively straight forward fashion. To demonstrate how this works, we'll consider baseball induced cheer (BIC) as our outcome, and attending a game at Fenway Park as our treatment. We want to see if attending a game at the Boston Red Sox stadium makes it more or less likely you'll experience postseason disappointment.

As a first pass, we look at the difference between the proportion of people who have been to a game at Fenway Park and experience BIC and the proportion of people who have

not been to a game at Fenway Park and experience BIC. We're assuming that the only cause of BIC is attending a game at Fenway Park, shown in the right side of Figure 6.1. This diagram also shows that this set up matches the most basic model for our chest pain patient, where hospitalization is the only cause of whether or not they revisit the emergency room. The (made up) data we'll be using is in Table 6.1.

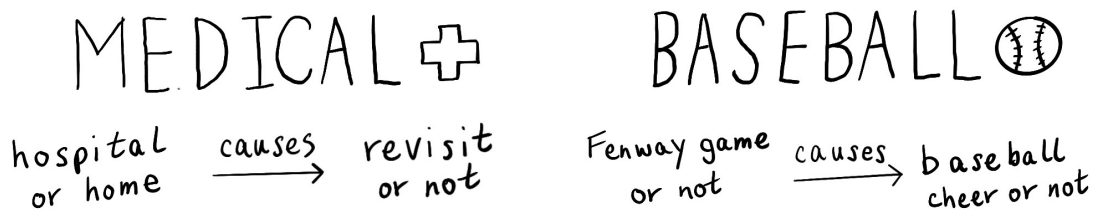

Figure 6.1: Our first model. For the medical example, we assume the only cause of emergency department revisits is the choice of hospitalization or discharge home. For the baseball example, we assume the only cause of baseball cheer (BIC) is whether or not you went to a Fenway Park game.

	Had BIC	Did not have BIC	Total
Went to a game	870	1430	2300
Did not go to a game	270	2430	2700
Total	1140	3860	5000

Table 6.1: Total counts of people categorized by whether or not they went to a game at Fenway Park and experienced BIC.

Starting with the people who have been to a game, we see that 870/2300 had BIC. Of the people who did not go to a game, 270/2700 had BIC. Subtracting the two gives

$$870/2300 - 270/2700 \approx 0.28$$

(The $\approx$ symbol means "approximately" since the fractions didn't quite work out neatly this time. For example, we could write $1.2 \approx 1$ since 1.2 is approximately 1). This is our estimate without considering confounding. Since it is positive, this means going to a game at Fenway increases the chance you get BIC.

Our confounder will be whether or not someone lives in New England because New

Englander's are more likely to be Red Sox's fans and live closer to the stadium than the average American, which means New Englander status is likely to have an effect on both postseason cheer and whether or not someone has been to a game. This is shown in the right side of Figure 6.2. To handle this, we separately estimate the effect for New Englanders and non-New Englanders, and then combine back together weighted by the proportion of people who are in each category.

	New Englander		non-New Englander	
	Had BIC	Did not	Had BIC	Did not
Went to a game	560	240	310	1190
Did not	0	200	270	2430
	1000		4000	

Table 6.2: Total counts of people categorized by whether or not they went to a game at Fenway Park, had BIC, and are New Englanders.

Looking at only the New Englanders in Table 6.2, we see that the proportion of people who went to a game and had BIC was 560/800, and the proportion of people who did not go to a game but had BIC is 0/200, so the estimated effect for New Englanders is

$$560/800 - 0/200 = 0.7.$$

This means that for New Englanders, going to a baseball game at Fenway greatly increases the probability that you experience baseball induced cheer. For non-New Englanders, the proportion who went to a gave and had BIC is 310/1500 and the proportion who did not go to a game but had BIC is 270/2700, so the estimated effect for non-New Englanders is

$$310/1500 - 270/2700 \approx 0.11.$$

This is a much smaller increase than we saw for New Englanders, which makes sense since New Englanders are more likely to be Red Sox fans, and going to a game at Fenway park is probably more fun for Red Sox fans.

To combine these two estimates, we note that 1000 out of the total 5000 people are

New Englanders and 4000 are not. We then use those fractions as weights, so that the estimate for New Englanders has less influence than the non-New Englander estimate since most people are not New Englanders, so we get

$$\frac{1000}{5000}(0.7) + \frac{4000}{5000}(0.11) \approx 0.22.$$

This is called direct adjustment, because we're directly adjusting for the effect of being a New Englander on our estimate. As you can see, our estimate decreased from 0.28 the first way we computed it to 0.22 the second way, meaning that going to a baseball game at Fenway has less of an effect on the chances that someone experiences baseball joy than we originally thought.

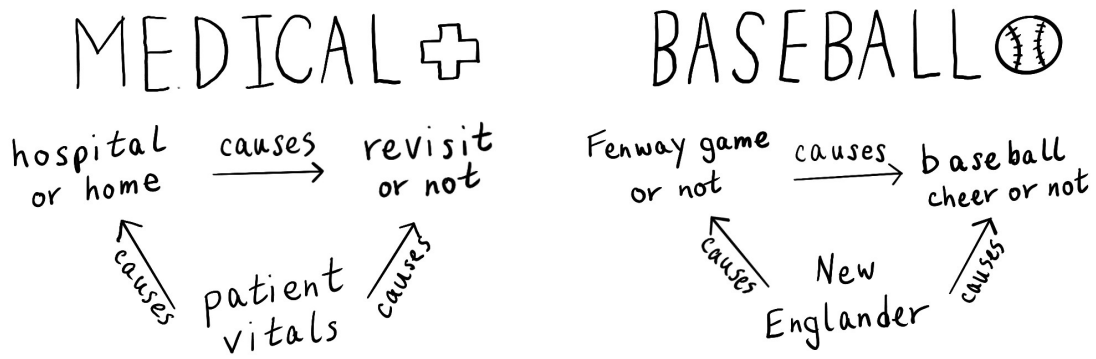

Figure 6.2: Our adjustment model. For the medical model, we assume that the patient's vital signs are a cause for both the patient's treatment and if they revisit. For the baseball model, we assume that whether or not someone is a New Englander is a cause for both if they go to a Fenway game and if they experience baseball cheer.

Returning to healthcare research, in practice we often adjust for as many "pre-treatment variables" (variables we can measure before we assign a treatment) as possible. For the chest pain study this would include vital signs like heart rate, blood pressure, age, sex, and body temperature. As shown in Figure 6.2, patient vitals and New Englander status play similar rolls. We account for patient vitals, which makes sure that we're comparing outcomes for similar patients, rather than comparing outcomes for potentially very different patients. Otherwise, we are assuming the effect is the same for everybody, which could

cause problems. If a drug works well for most people, but not for a specific subgroup, we could end up with an overly optimistic estimate, just like our first estimate of 0.28 was overly optimistic compared to our more careful estimate of 0.22 that incorporated subgroups (New Englanders and non-New Englanders). Taking a weighted averaging of all the different subgroup effects gives us a more realistic estimate.

Up until now, we've been discussing standard practice. We adjust for as much as possible, using the information that we have. But what happens if we don't have all the information we want? In the baseball example, we're comparing New Englanders to New Englanders and non-New Englanders to non-New Englanders, assuming that the effect of a Red Sox game is going to be different for those two groups. This is the commonly used approach to dealing with confounding. However, I would argue that we're really using New Englander status as a way to guess at whether or not someone is a Red Sox fan. That means that Red Sox fan status is our actual confounder, but we don't have access to that in our data set so we're using New Englander status to make an educated guess. This happens in medical records all the time, where we don't have all the data we would ideally use (because medical records are designed for wide practical use, not tailored to specific research projects), so I developed a new method that assumes we can't use the confounder directly, but we have what I will call *proxies*, imperfect measurements of the actual confounders.

6.3 Method

Let's suppose that we go back to our data and realize that we also have information on who watches softball/baseball. This, in combination with our New Englander information, gives us Table 6.3.

	New Englander	non-New Englander
Watches softball/baseball	1000	1000
Does not	400	2600

Table 6.3: Total counts of people categorized by whether or not they watch softball/baseball and are New Englanders.

I would argue that New Englander status and softball/baseball watching status do not directly affect if someone goes to a game at Fenway Park or if they experience BIC, but instead tell us information about whether or not someone is a Red Sox fan. Moving forward I'll call this type of variable a *proxy*, because it is not actually the information we want (if someone is a Red Sox fan) but it acts like a replacement for that information. If someone is a Red Sox fan, a different team fan (non-Red Sox fan), or not a baseball fan (non-fan) is a latent variable, meaning something that we cannot directly measure or don't have access to. In Figure 6.3, these latent variables are circled in red.

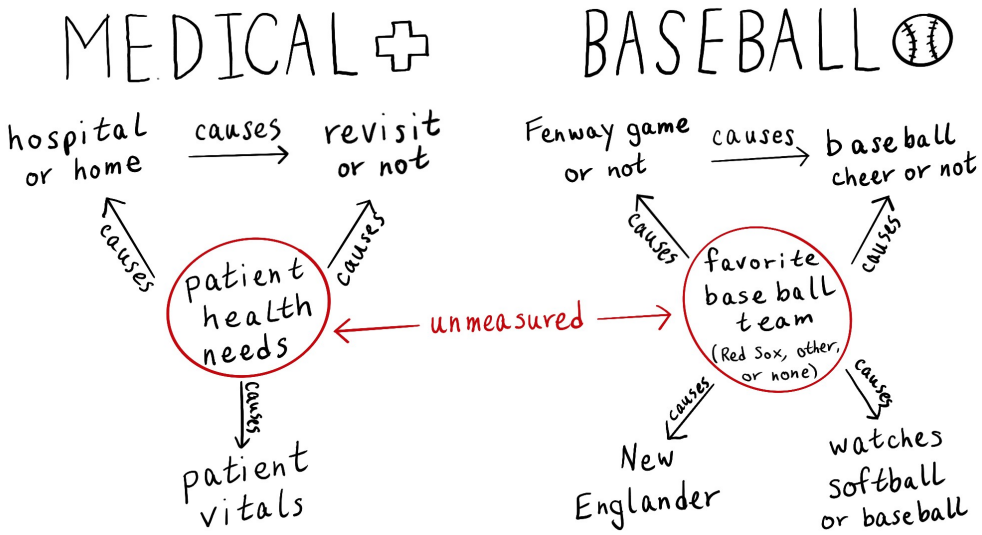

Figure 6.3: Our latent variable model. For the medical example, we assume that a patient's vital signs give us information about their underlying needs, which affects both their treatment and if they revisit. For the baseball example, we assume that if they're a New Englander and if they watch softball/baseball gives us information about their favorite baseball team, which affects both if they go to a Fenway game and if the experience baseball cheer. The latent variables, the things we cannot directly measure or don't have information on, are in red circles.

I know it feels strange that we would know if someone went to a baseball game and if they watch softball or baseball but not if they're fans of a specific team. The oddities of pre-existing data sets make this kind of thing happen all the time. One database I worked with included if someone had insurance, and what medication they were prescribed, but did not include which kind of insurance so we couldn't tell if their treatment was what

the doctor originally wanted or what they settled on because it was what insurance would cover. We make do with what we can find.

It would be reasonable to guess that the 1000 people who are both New Englanders and softball/baseball watchers are probably Red Sox fans. The 1000 people who watch softball/baseball but are not New Englanders are most likely baseball fans but of a different team. The 3000 people who do not watch softball/baseball probably are not baseball fans regardless of where they live. A person's status as a baseball fan or not and a Red Sox fan or not are not directly available to us but we can use New Englander status and softball/baseball watching as proxies to approximate for fan status. Not every baseball watching New Englander is a Red Sox fan, but most will be, and being a Red Sox fan is far more likely to influence baseball induced cheer directly than whether or not you're a New Englander.

With this in mind, we once again recategorize our data in Table 6.4, assuming that softball/baseball watching New Englanders are one group, softball/baseball watching non-New Englanders are one group, and people who don't watch softball/baseball are one group.

	Probable Red Sox fans		Probable other fans		Probable non-fans	
	Had BIC	Did not	Had BIC	Did not	Had BIC	Did not
Went to a game	720	80	50	450	100	900
Did not	120	80	150	350	0	2000
	1000		1000		3000	

Table 6.4: Total counts of people categorized by whether or not they went to a game at Fenway Park, and had BIC, in addition to our approximation of who is a probable Red Sox fan, probable other team fan, or probable non-baseball fan.

Instead of grouping by something directly measured in our data set, we're grouping by our best guess at the latent variable of being a Red Sox fan, a fan of a different team, or not a baseball fan, but we'll use the same weighting trick we did with the New Englander adjusted estimate.

For probable Red Sox fans: $\frac{720}{800} - \frac{120}{200} = 0.3$.

For probable other team baseball fans: $\frac{50}{500} - \frac{150}{350} = -0.2$.

For probable non-baseball fans: $\frac{100}{900} - \frac{0}{2000} = 0.1$.

Combining them all proportionally gives

$$\frac{1000}{5000}(0.3) + \frac{1000}{5000}(-0.2) + \frac{3000}{5000}(0.1) = 0.08.$$

This is even lower of an estimate than we got last time! A 0.08 increase in the rate of baseball induced cheer for going to a game at Fenway Park is very different from a 0.28 increase. Figure 6.4 shows that at each step our overall estimate (the red box) just kept getting lower. How do we choose which estimate to use in practice? We talk to subject experts about whether being a New Englander or not, or being a Red Sox fan, a non-Red Sox fan, or a non-baseball fan is more informative on both if someone will go to a game at Fenway and if they'll experience BIC.

This process of combining multiple proxies of some underlying latent variable (remember, latent means it's not measured in our data) to get better categorizations is the basis for the method I proposed and tested for use with medical data. For example, (and take this with a grain of salt, I'm hopefully a math doctor but I'm definitely not a medical doctor) your blood pressure may not directly affect whether or not you come back to the emergency department within a month. Blood pressure is, however, an indicator of your cardiac health, which does directly affect whether or not you come back to the emergency department within a month, but is not directly measurable. In this more general method, we do some modeling to approximate the general behavior of the latent variable (e.g. cardiac health) using the observed information (e.g. blood pressure and heart rate), use the information we get from that first model to set up a second model to estimate the effect for each group, and then combine everything proportionally to find the effect.

One strength of this method is that we can get estimates for subgroups. Figure 6.4 shows how different the effect is for the different subgroups in our baseball example. For probable Red Sox fans, meaning New Englanders who watch softball/baseball, our final estimate was 0.3. This is higher than any of our overall estimates. For this subgroup of people, going to a game at Fenway greatly increases your chance of baseball induced cheer.

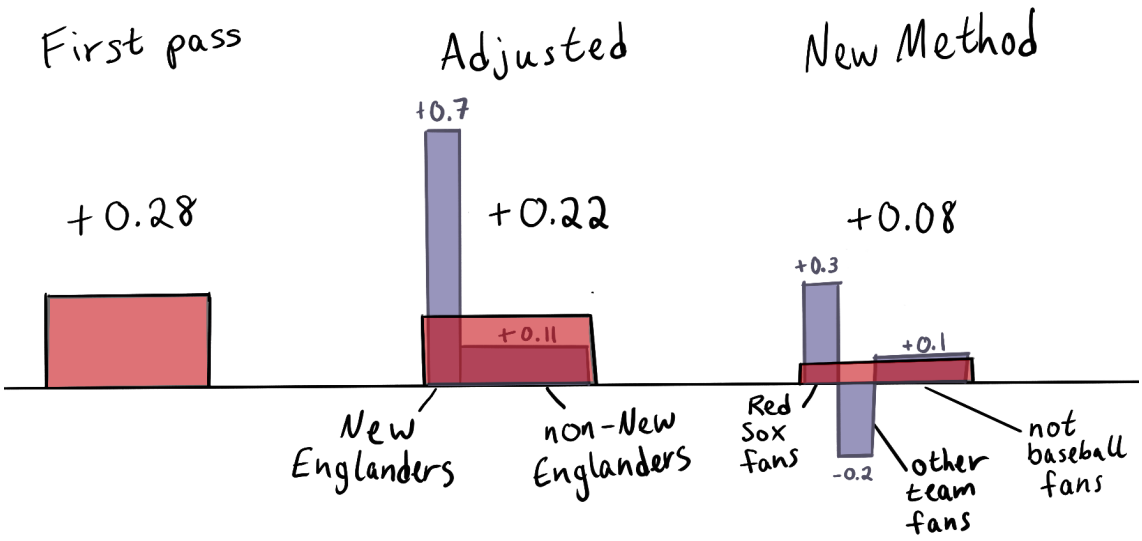

Figure 6.4: The estimates from each of our different methods. The purple boxes show the subgroup estimates that are combined to get the overall estimates, shown in red. We can see how much the effect differs by subgroup, especially with the new method.

Meanwhile, for softball/baseball watching non-New Englanders (fans of other teams), our final estimate was -0.2! For this subgroup, going to Fenway decreases the chance that you enjoy baseball! These subgroup level effect estimates can be very useful in medicine. What helps one type of patient might hurt another, so more targeted subgroup estimates can lead to more tailored treatment decisions that help patients.

I spent a lot of time staring at chalk boards trying to sort out how to properly link the two models without knowing beforehand what the exact models will be, so that people using my method can use the models that fit their situation. I wanted to create a framework that would be as flexible as possible but still work. One way to think about this is to imagine you're trying to design something like Legos from scratch. At this point there are hundreds of different Lego bricks that can all be used together and combined in more ways than I can imagine. This works because of the framework they are developed within. There's specific heights and widths each piece has to be to make sure they all fit together, and each piece has to have either studs (the round bumps that stick out) or tubes (where the studs fit in to) so they can be combined with other pieces. These requirements make

sure all the pieces work together, but does not specify exactly what each brick needs to be.

I needed to figure out what needed to be true about the models (for Lego this is the allowable sizes) and how to fit them together (the studs and tubes system of connections). I would pick some models, and then start at the top of the board with the information this gave me, statements like "The model says if I know the confounder, then the proxies don't give me any extra information about the outcome, so the model using both the proxies and the confounder is the same as the model using just the confounder" or "The model says I only know the average of the confounder for each subgroup, not each individual value." These are very wordy written out like this, but much more concise when written symbolically so that is what I did. At the bottom of the board would go the thing I was trying to estimate, the causal effect.

I spent ages using the rules of algebra and probability to rearrange the symbols over and over, trying to get the top of the board to match the bottom, or working backwards trying to get the bottom to match the top. This is like trying to plan out a road trip, where you know where you start and where you want to end up, but need to connect everything in the middle. You can't always drive the most convenient direction because you have to follow the roads. I knew the information I was starting with, and the information I wanted to find, but I needed to figure out the right steps in the middle, following the rules of mathematics.

As I worked, I tried adding different assumptions to open up extra paths I could use in my rearranging. Going back to Legos, if you're trying to plan out how to combine a handful of bricks but don't put any restrictions on what those bricks could be, there's not a lot you can be sure you'll be able to do. What happens if your plan says to stack two bricks on top of each other but when you do finally get handed a specific pair of bricks they happen to both be flat on top? Figure 6.5 gives an example of when this can't work out. If you add the assumption that one brick has tubes on the bottom and the other has studs on top, that narrows down the possible pairs of bricks but allows you to add "stack

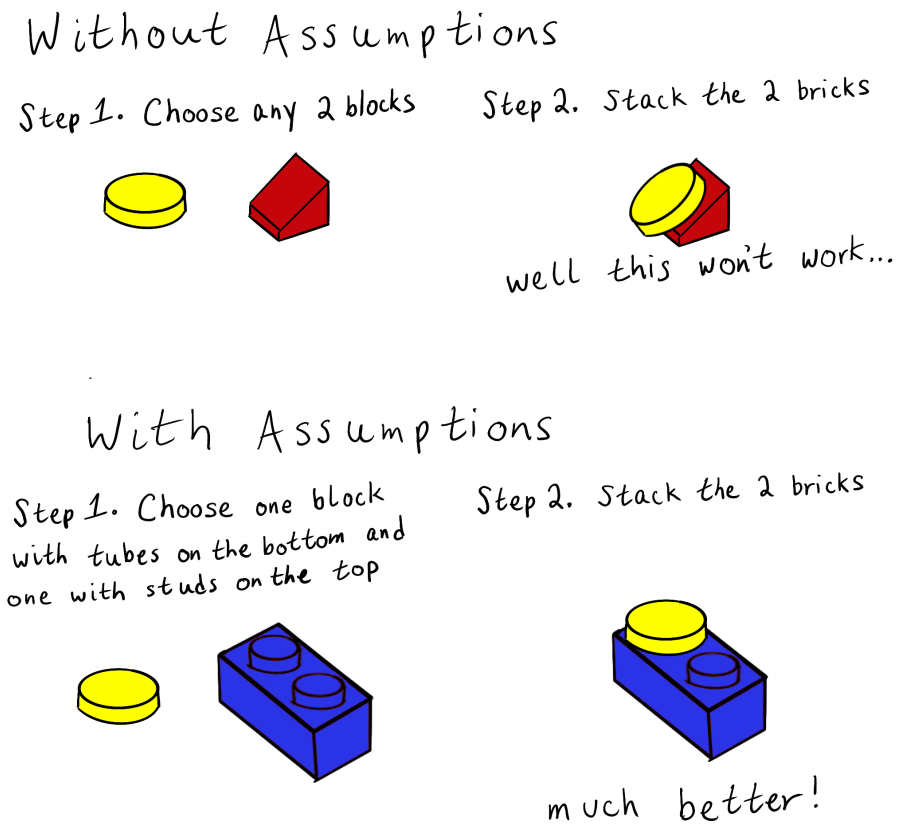

Figure 6.5: An example showing that sometimes, to do what we want to do, we need to add some assumptions, or limitations.

these two bricks" to your plan. In math we have the same problem. Ideally every move we make would always work, but when that is not possible (and it's almost always not possible) we add some restrictions so that we can make progress.

Based on trying a lot of combinations of models, I eventually settled on two main assumptions: 1. the proxies and the treatment give us enough information to estimate the model for the unmeasured confounder, and 2. the causal effect for each group can be written as the product of the effect from the treatment and the effect from the unmeasured confounder (if we take the effect from the treatment and the effect from the confounder and multiply them together we get the effect). As long as the two models meet those requirements, we can fit them together to get our estimate. We learn what we can about

the unmeasured confounder, use that to estimate the treatment effect part, and then do a fancy version of taking a weighted average that works even if you have infinitely many subgroups (like if your subgroups are decided by breathing rate, so 15 breaths per minute is one group, 15.1 breaths per minute is another group, and so on). This plan works for a lot of different models, creating a unified framework that is flexible and works for a variety of research studies.

6.4 Conclusion

Along with a detailed mathematical argument for why my method should work in theory, I also needed some proof that it works in practice. Since with real data we almost never know the true effect, I designed a set of different simulations where I knew what the right answer should be, and used my method and the usual methods to estimate the effect. Each simulation tested something different, including what happens for different sample sizes or if some of the assumptions are wrong. My method worked surprisingly well in many of these simulations, but the tests where it did poorly show where further research should focus in the future.

This method is an improvement, but it is not perfect. You might have noticed that when we labeled something a proxy, we made the assumption that it does not directly influence the treatment (going to a game at Fenway) or the outcome (baseball induced cheer). This assumption is practically impossible. Whether or not you live in New England influences how far you need to travel to get to a game, thereby impacting your chances of going to a game, and also influences how stressful the travel will be which impacts your cheer chances.

In a healthcare setting, this would translate to situations like assuming that blood pressure is a proxy for cardiac health, and does not directly impact either the treatment your healthcare provider chooses for you, or your outcome. On the one hand, it is easy to argue that your cardiac health, this underlying thing we cannot directly measure, is really what the provider is trying to improve, and what determines your outcome. On the other

hand, high blood pressure is often directly treated in the emergency department, and can greatly increase the risk of stroke and other forms of organ damage [108]. It could be either a proxy or a confounder, depending on the context of the research.

Right now, it would be a judgment call on the part of the researcher to determine in their specific setting whether it is more accurate to treat something as a proxy or as a confounder because we do not have methods that will give you a good answer if you categorize incorrectly. In my case study of hospitalization for older adults in the emergency department with chest pain, I chose to treat blood pressure as a proxy but that was a decision I had to make because my method treats confounders and proxies differently. I have to rely on subject experts, and make the choice based on imperfect information. This is a shortcoming of my method.

A common saying in statistics is "All models are wrong, but some are useful." To get anywhere, we have to make some assumptions about the world, the data, and how all the pieces relate to each other. It is useful that my method does not assume we have direct access to the confounder, since that is a common problem in research using observational data sets that were not tailored to the particular project. It is probably wrong to assume that we can always correctly decide if something is a proxy or a confounder. I am hoping that I can find a way to combine my proxy method with another confounder method (like direct adjustment) to create something that does not require either assumption, so it can be used in even more situations. This is the usual cycle of math, we make imperfect progress built on imperfect progress.

In the meantime, I'm hoping that medical researchers can use my method in their projects. I want to create easy to read documentation for how it works, even more examples of it in action, and potentially even a software package that makes it easy for medical professionals to use. I want to help practitioners make better decisions, and choose better treatments, for all the people they take care of every day.