WEBVTT

00:00:00.000 –> 00:00:02.400 - Maybe one or two minutes and then,
00:00:02.400 –> 00:00:03.233 I’ll have you introduced.
00:00:03.233 –> 00:00:04.640 - And it’s about, and so I…
00:00:04.640 –> 00:00:06.860 And it’s gonna be more fun for me if it’s a little
00:00:06.860 –> 00:00:08.510 interactive, as much as we can make it.
00:00:08.510 –> 00:00:11.760 So I won’t be able to see all of you nodding and
whatnot,
00:00:11.760 –> 00:00:14.827 but please feel free to jump in.
00:00:14.827 –> 00:00:16.830 And the talk’s gonna be pretty non-technical.
00:00:16.830 –> 00:00:18.960 My goal is mostly to sort of help
00:00:18.960 –> 00:00:23.360 convey some of the concepts and ideas and so I
will.
00:00:23.360 –> 00:00:27.143 Hopefully it will be a reasonable topic to do via
Zoom.
00:00:30.050 –> 00:00:31.420 Great, so I think,
00:00:32.650 –> 00:00:35.670 Frank basically gave this stuff that’s relevant
00:00:35.670 –> 00:00:38.890 on this slide.
00:00:36.800 –> 00:00:38.900 I do also wanna apologize, those of you guys
00:00:38.890 –> 00:00:41.170 who I was supposed to meet with this morning,
we have a...
00:00:41.170 –> 00:00:43.850 My husband broke his collarbone over the weekend.
00:00:43.850 –> 00:00:46.840 So I’ve had to cancel things this morning,
00:00:46.840 –> 00:00:49.853 but I’m glad I’m able to still do this seminar,
00:00:51.278 –> 00:00:52.380 I didn’t wanna,
00:00:52.380 –> 00:00:53.380 have to cancel that.
00:00:54.350 –> 00:00:56.060 So again,
00:00:56.060 –> 00:00:58.530 the topic is gonna be sort of this idea of external
00:00:58.530 –> 00:01:01.330 validity, which I think is a topic that people often
00:01:01.330 –> 00:01:03.540 are interested in because it’s the sort of thing
00:01:03.540 –> 00:01:06.170 that we often think sort of qualitatively about,
but there hasn’t been a lot of work thinking about it quantitatively. So again, my goal today will be to sort of help give a framework for thinking about external validity in sort of a more formal way. So let’s start out with the sorts of questions that might be relevant when you’re thinking about external validity. So it might be research questions like a health insurer is deciding whether or not to approve some new treatment for back pain. There might be interested predicting overall population impacts of a broad public health media campaign. A physician practice might be deciding whether training providers in a new intervention would actually be cost effective given the patient population that they have. And that I felt like I needed to get some COVID example in... But, for example, a healthcare system, might wanna know whether it’s sort of giving convalescent plasma to all of the individuals recently diagnosed with COVID-19 in their system, whether that would sort of lead to better outcomes overall. So all of these...
What I’m distinguishing here or sort of trying to convey is that all of these reflect what I will call a population average treatment effect. So across some well-defined population, does some intervention work sort of on average. The population might be pretty narrow. Again, it might be the patients in one particular physician practice, or might be quite broad. It could be everyone in the State of Connecticut or in the entire country. But either way, it’s a well-defined kind of population and we’ll come back to that.

What’s really important, and this will sort of underlie much of the talk is that kind of the whole point is that there might be underlying treatment effect heterogeneity. So there might be some individuals for whom this treatment of interest is actually more effective than others. But what I wanna be clear about, is the goal of inference that I’m talking about today, is gonna be about this overall population average. So we’re not trying to say like which people are gonna benefit more or sort of to which people should we give this treatment. It’s really more a question of sort of more population level decisions, sort of if we have...
If we’re making a decision, that’s sort of a policy kind of population level, on average is this gonna be something that makes sense. So I hope that distinction makes sense. I’m happy to come back to that. So again until I don’t know, five or, well maybe now more than 10 years ago, there had been relatively little attention to the question of how well results from kind of well-designed studies like a randomized trial might carry over to a relevant target population. I think in much of statistics as well as fields like education research, public policy, even healthcare, there’s really been a focus on randomized trials and getting internal validity, and I’ll formalize this in a minute. But in the past 10 or so years, there’s been more and more interest in this idea of how well can we take the results from a particular study and then project them to well-defined target population. And again, so today I’m gonna try to give sort of an overview of the thinking in this area, along with some of the limitations and in particular, the data limitations that we have in thinking about this. One thing I do wanna be clear about is there’s a lot
00:04:35.720 –> 00:04:38.010 of reasons why results from randomized trials
00:04:38.010 –> 00:04:39.580 might not generalize.
00:04:39.580 –> 00:04:42.320 There’s some classic examples in education
00:04:42.320 –> 00:04:44.450 where there are scale-up problems.
00:04:44.450 –> 00:04:47.903 The classic example is one I’m looking at,
00:04:49.890 –> 00:04:50.750 class size.
00:04:50.750 –> 00:04:53.880 And so, in Tennessee, they randomly assign kids
00:04:53.880 –> 00:04:56.620 to be in smaller versus larger classes
00:04:56.620 –> 00:04:59.570 and found quite large effects of smaller classes.
00:04:59.570 –> 00:05:02.530 But then, when the State of California tried to
00:05:02.530 –> 00:05:05.880 implement
00:05:05.880 –> 00:05:08.040 this, the problem is that you need a lot more
00:05:08.040 –> 00:05:10.720 teachers
00:05:10.720 –> 00:05:11.553 to kind of roll that out statewide.
00:05:11.553 –> 00:05:13.970 And so, it led actually to a different pool of teach-
00:05:13.970 –> 00:05:16.170 ers
00:05:16.170 –> 00:05:19.010 to different contexts or different implementation.
00:05:19.010 –> 00:05:21.250 Today, what I’m gonna be focusing on are differ-
00:05:21.250 –> 00:05:23.503ences
00:05:23.503 –> 00:05:25.770 between a sample and a population.
00:05:25.770 –> 00:05:27.630 Their difference is in sort of baseline characteris-
00:05:27.630 –> 00:05:28.757 tics,
00:05:28.757 –> 00:05:31.763 that moderate treatment effects.
00:05:31.763 –> 00:05:34.230 And again, I’ll formalize this a little bit as we go
00:05:34.230 –> 00:05:36.830 along.
00:05:36.830 –> 00:05:38.740 Just as a little bit of an aside,
00:05:38.740 –> 00:05:40.000 in case some of you know this field a little bit,
00:05:40.000 –> 00:05:42.320 just to give you a little, just...
Some people might use the term transportability. So some of the literature in this field uses the term transportability. I tend to use generalizability. There’s some subtle differences between the two, which we can come back to, but for all intents and purposes, like they basically can think of them interchangeably for now. I also wanna note, if any of you kind of come from like a survey world, these debates about kind of how well a particular sample reflects a target population are exactly, not exactly the same, but very similar to the debates happening in the survey world. Andrew Mercer has a nice summary of that. Again, I’m happy to talk more about that. Okay, any questions before I keep going? Okay.
So let me formalize kind of what we’re talking about.

This is…

This framework is now, 12 years old.

Time goes quickly.

But we’re just to formalize what we’re interested in.

The goal is to estimate, again, this what I’ll call a population average treatment effect or PATE.

And so here,

hopefully you’re familiar with sort of potential outcomes

and causal inference.

But the idea is that we have some well-defined population

of size N.

And Y(1) is the potential outcomes, if people in that population receive the treatment condition

of interest.

Y(0) are the outcomes if they receive the control or comparison condition of interest.

So here, we’re just saying we’re interested in the average effect, basically sort of the difference in potential outcomes, average across the population.

We could be doing this with risk ratios

or odds ratios or something.

Those are a little more complicated because the math doesn’t work as nicely.

So for now think about it more like risk differences

or something, if you have a binary outcome,
the same fundamental points hold. So I’m not gonna tell you right now where the data we have came from, but imagine that we just have a simple estimate of this PATE, as the difference in means of some outcome between an observed treated group and an observed control group. So again, we see that there’s a bunch of people who got treated, a bunch of people who got control, and we might estimate this PATE as just the simple difference in means between again, the treatment group and the control group. So what I wanna talk through for the next couple of minutes, is the bias in this sort of naive estimate of the PATE. So we’ll call that Delta. So I’m being a little loose with notation here, but sort of the PATE that the bias essentially think of it as sort of the difference between the true population effect and our naive estimate of it. And what this paper did with Gary King and Kosuke Imai, we sort of laid how different choices of study designs impact the size of this bias. And in particular, we showed that sort of under some simplifying situations, sort of mathematical simplicity,
you can decompose that overall bias into four pieces.

So the two Delta S terms are what are called, what we call sample selection bias.

So basically, the bias that comes in if our data sample is not representative of the target population that we care about.

The Delta T terms are our typical sort of confounding bias.

The X refers to the variables we observe, and the U refers to variables that we don’t observe.

So what we then did in the paper, and this is sort of what motivates a lot of this work is to think through these, again, the trade offs in these different designs.

And essentially what we’re trying to sort of point out is that...

Let’s go to the second row of this table first actually, a typical experiment.

So a typical experiment, I would say is one where we kind of take whoever comes in the door, we kind of try to recruit people for a randomized trial, whether that’s schools or patients or whatever it is.

And we randomized them to treatment and control groups.
So that is our typical randomized experiment. The treatment selection bias in that case is zero. In expectation, that’s why we like randomized experiments. In expectation, there is no confounding. and we get an unbiased treatment effect estimate for the sample at hand. The problem for population inference is that the Delta S terms might be big, because the people that agree to be in a randomized trial, might be quite different from the overall population that we care about. So in this paper, we’re trying to just sort of… In some ways, be a little provocative and point this out that our standard thinking about study designs and sort of our prioritization of randomized trials, implicitly prioritizes internal validity over external validity. And in particular, if we really care about population effects, we really should be thinking about these together and trying to sort of have small sample selection bias and small treatment selection bias. So an ideal experiment would be one where we can randomly select people for our trial. Let’s say we have… Well, actually, I’ll come back to that in a second.
Randomly select people for our trial and then randomly assign people to treatment or control groups. And in expectation, we will have zero bias in our population effect estimate.

But these other designs, and again, like a typical experiment might end up having larger bias overall, than a well designed non-experimental study, where if we do a really good job like adjusting for confounders, it may be that well done non-experimental study conducted using say the electronic health records from a healthcare system might actually give us lower bias for a population effect estimate.

Then does a non-representative small randomized trial. Again, a little provocative, but I think useful to be thinking about what is really our target of inference and how do we get data that is most relevant for that.

I will also just as a small aside, maybe a little on the personal side, it’s been striking to me in the past two days. So my husband broke his collarbone over the weekend. And it turns out the break is one where there’s a little bit
00:12:34.730 –> 00:12:37.760 of debate about whether you should have surgery or not.
00:12:37.760 –> 00:12:39.360 Although kind of recent thinking is that
00:12:39.360 –> 00:12:40.290 there should be surgery.
00:12:40.290 –> 00:12:44.240 And I was doing a PubMed search as a good statistician
00:12:44.240 –> 00:12:46.970 public health person whose family member
00:12:46.970 –> 00:12:49.300 needs medical treatment.
00:12:49.300 –> 00:12:51.790 And I found all these randomized trials that actually
00:12:51.790 –> 00:12:54.910 randomized people to get surgery or not.
00:12:54.910 –> 00:12:56.000 And then I came home...
00:12:56.000 –> 00:13:00.320 Oh, no, I didn’t come home, we were home all the time.
00:12:58.750 –> 00:13:03.280 I asked my husband later, I was like,
00:13:03.280 –> 00:13:06.770 would you ever agree to be randomized?
00:13:06.770 –> 00:13:09.050 And would we ever agree to be randomized?
00:13:09.050 –> 00:13:11.065 And he’s like, no, we wouldn’t.
00:13:11.065 –> 00:13:14.550 We’re gonna go with what the physician recommends
00:13:14.550 –> 00:13:16.300 and what we feel is comfortable.
00:13:16.300 –> 00:13:19.250 And it really just hit home for me at this point that
00:13:19.250 –> 00:13:22.070 the people who agree to be randomized or the context
00:13:22.070 –> 00:13:25.860 under which we can sort of randomize
00:13:25.860 –> 00:13:27.730 are sometimes fairly limited.
00:13:27.730 –> 00:13:31.230 And again, so partly what this body of research is trying
00:13:31.230 –> 00:13:33.410 to do is sort of think through what are the implications
of that when we do wanna make population inferences.

Make sense so far?

I can’t see faces, so hopefully.

Okay.

So, I will say a lot of my work in this area has actually, in part been just helping or trying to raise awareness of thinking about external validity bias.

So some of the research in this area has been trying to understand how big of a problem is this.

If maybe people don’t agree to be in randomized trials very often,

but maybe that doesn’t really cause bias in terms of our population effect estimates.

So what I’ve done in a couple of papers on these other sides on this slide is basically trying to formalize this and it’s pretty intuitive, but basically we show, and I’m not showing you the formulas here.

But intuitively, there will be bias in a population effect estimate essentially if participation in the trial is associated with the size of the impacts.

So in particular, what I’ll call the external validity bias.

So, those Delta S terms kind of the bias due to the lack of representativeness is a function of the variation of the probabilities.
of participating in a trial, variation and treatment effects, and then the correlation between those things. So if constant...

If we have treat constant treatment effects or the treatment effect is zero or is two for everyone, there's gonna be no external validity bias. It doesn't matter who is in our study.

Or if everyone has an equal probability of participating in the study, we really do have a nice random selection, then again, there's gonna be no external validity bias.

Or if the factors that influence whether or not you participate in the study are independent of the factors that moderate treatment effects, again, there'll be no external validity bias.

The problem is that we often have very limited information about these pieces. We, as a field, I think medicine, public health, education, all the fields I worked in, there has not been much attention paid to these processes of how we actually enroll people in studies.

And so it’s hard to know kind of what factors relate to those and if those then also moderate treatment effects.
00:15:53.064 –> 00:15:54.103 (phone ringing)
00:15:54.103 –> 00:15:55.360 Oops, sorry.
00:15:55.360 –> 00:15:57.800 Incoming phone call, which I will ignore.
00:15:57.800 –> 00:15:58.890 So,
00:15:58.890 –> 00:16:00.100 there has been...
00:16:01.010 –> 00:16:01.843 Sorry.
00:16:02.950 –> 00:16:05.310 There has been a little bit of work trying to doc-
00:16:05.310 –> 00:16:10.310 in real data and find empirical evidence on these
sizes.
00:16:10.780 –> 00:16:13.000 The problem, and sorry, some of the...
00:16:13.000 –> 00:16:13.950 Some of you might...
00:16:13.950 –> 00:16:15.820 If any of you are familiar with the, like,
00:16:15.820 –> 00:16:18.230 within what it’s called the within study comparison
00:16:18.230 –> 00:16:19.063 literature.
00:16:19.063 –> 00:16:21.750 So there’s this whole literature on non-
00:16:23.240 –> 00:16:27.570 that sort of try to estimate the bias due to non-
random
00:16:27.570 –> 00:16:29.700 treatment assignment.
00:16:29.700 –> 00:16:31.510 This is sort of analogous to that.
00:16:31.510 –> 00:16:33.710 But the problem here is that what you need is you
need
00:16:33.710 –> 00:16:37.240 an accurate estimate of the impact in the popula-
00:16:37.240 –> 00:16:40.140 And then you also need sort of estimates of the
impact
00:16:40.140 –> 00:16:43.690 in samples that are sort of obtained in kind of
typical
00:16:43.690 –> 00:16:44.990 ways.
00:16:44.990 –> 00:16:46.690 So that’s actually really hard to do.
00:16:46.690 –> 00:16:49.050 So I’ll just briefly talk through two examples.
00:16:49.050 –> 00:16:51.810 And if any of you have data examples that you think might
00:16:51.810 –> 00:16:54.570 sort of be useful for generating evidence,
00:16:54.570 –> 00:16:56.800 that would be incredibly useful.
00:16:56.800 –> 00:16:58.880 So one of the examples is...
00:17:00.050 –> 00:17:01.750 So let me back up for a second.
00:17:01.750 –> 00:17:03.330 In the field of mental health research,
00:17:03.330 –> 00:17:05.530 there’s been a push recently, or actually not so much
00:17:05.530 –> 00:17:08.270 recently in the past, like 10, 15 years
00:17:08.270 –> 00:17:11.810 to do what I call or what are called pragmatic
00:17:11.810 –> 00:17:14.760 trials with the idea of enrolling much more...
00:17:15.910 –> 00:17:20.710 A much broader set of people use a broader set of
practices
00:17:20.710 –> 00:17:22.393 or locations around the country.
00:17:23.400 –> 00:17:26.620 And so what this Wisniewski et al people did was
they took
00:17:26.620 –> 00:17:28.940 the data from one of those large pragmatic trials.
00:17:28.940 –> 00:17:29.773 And the idea they...
00:17:29.773 –> 00:17:32.530 Again, the idea was that it should be more repre-
sentative
00:17:32.530 –> 00:17:35.070 of people in this case with depression
00:17:35.070 –> 00:17:36.830 across the U.S.
00:17:36.830 –> 00:17:38.100 And then, they said, well, what if...
00:17:38.100 –> 00:17:39.560 In fact, we didn’t have that.
00:17:39.560 –> 00:17:43.760 What if we use sort of our normal study inclusion
00:17:43.760 –> 00:17:47.360 and exclusion criteria, it’s sort of been, we’d like
subset,
00:17:47.360 –> 00:17:49.960 this pragmatic trial data to the people that we
think
00:17:49.960 –> 00:17:53.260 would have been more typically included in a sort
of
more
00:17:53.260 –> 00:17:55.220 standard randomized trial.
00:17:55.220 –> 00:17:57.740 And sort of not surprisingly, they found that
the people in the sort of what they call the efficacy sample, those sort of typical trial sample had better outcomes and larger treatment effects than the overall pragmatic trial sample as a whole. We did something similar sort of in education research where it’s a little bit in the weeds. I don’t really wanna get into the details, but we essentially had a pretty reasonable regression discontinuity design. So we were able to get estimates of the effects of this reading first intervention across a number of states. And we then compared those state wide impact estimates to the estimates you would get if we enrolled only the sorts of schools and school districts that are typically included in educational evaluations. And there we found that this external validity bias was about 0.1 standard deviations, which in education world is fairly large. Certainly people would be concerned about an internal validity bias of that size. So we were able to sort of use this to say, look, if we really wanna be serious about external validity, it might be as much of a problem as sort of typical internal validity bias that people care about in that field. So again, the problem though, is we don’t usually
have these sorts of designs where we have a population effect estimate, and then sample estimates, and we can compare them. And so instead we can sometimes try to get evidence on sort of the pieces. So, but again, we basically often have very little information on why people end up participating in trials. And we also are having, I think there’s growing numbers of methods, but there’s still limited information on treatment effect heterogeneity. Individual randomized trials are almost never powered to detect subgroup effects. Although, there is really growing research in this field and that is maybe a topic for another day. Okay. But again, there is a little... I think I’ll go through this really quickly, but, I will give credit to some fields which are trying to better understand kind of who are the people that enroll in trials and how do they compare policy populations of interest. So a lot of that has been done in sort of the substance use field. And you can see a bunch of sites here.
documenting that people who participate in randomized trials of substance use treatment do actually differ quite substantially from people seeking treatment for substance use problems more generally. So for example, the Okuda reference the eligibility criteria in cannabis treatment RCTs would exclude about 80% of patients across the U.S. seeking treatment for cannabis use. And so again, it’s sort of there’s indications that the people that participate in trials are not necessarily reflective of the people for whom decisions are having to be made. Okay, so hopefully that at least kind of give some motivation for why we want to think more carefully about the population average treatment effect and why we might wanna think about designing studies or analyzing data in ways that help us estimate that. Any questions before I move to, how do we do that? Okay. I will end... I’m gonna hopefully end it at about 12:45, 12:50, so we’ll have time at the end, too. So, as a statistician, I feel obligated to say, and actually I have a quote on this at the very end of the talk.
If we wanna be serious about estimating something, it’s better to incorporate that through the design of our study, rather than trying to do it post talk at the end. So let’s talk briefly about how we can improve external validity through study or randomized trial design. So again, as I alluded to earlier with the sort of ideal experiment, an ideal scenario is one where we can randomly sample from a population and then randomly assign treatment and control conditions. Doing this will give us a formerly unbiased treatment effect estimate in the population of interest. This is wonderful. I know of about six examples of this type. Most of the examples I know of are actually a federal government programs where they are administered through like centers or sites. And the federal government was able to mandate participation in an evaluation. So classic example is the Head Start Impact Study, where they were able to randomly select headstart centers to participate. And then within each center,
they randomized kids to be able to get in off the wait list versus not. An upward bound evaluation had a very similar design.

It’s funny, I was... I gave a talk on this topic at Facebook and I was like, why is Facebook gonna care about this? Because you would think at a place like Facebook, they have their user sample, they should be able to do randomization within, like they should be able to pick users randomly and then do any sort of random assignment they want within that.

It turns out it’s more complicated than that, and so, they were interested in this topic, but I think that’s another sort of example where people should be thinking, could we do this? Like, in a health system. I can imagine Geisinger or something implement something in their electronic health record where it’s about messaging or something. And you could imagine actually picking people randomly to then randomize.

But again, that’s pretty rare. There’s an idea that’s called purpose of sampling.
And this goes back to like the 1960s or 70s and the idea is sort of picking subjects purposefully.

So one example here is like maybe we think that this intervention might look different or have different effects for large versus small school districts.

So in our study, we just make an effort to enroll both large and small districts.

This is sort of nice. It kind of gives you some variability in the types of people or subjects in the trial, but, it doesn’t have the formal representativeness and sort of the formal unbiasedness, like the random sampling I just talked about.

And then again, sort of similar is this idea and this push in many fields towards pragmatic or practical clinical trials, where the idea is just to sort of try to enroll like kind of more representative sample in sort of a hand wavy way like I’m doing now.

So not, it doesn’t have this sort of formal statistical underpinning, but at least it’s trying to make sure that it’s not just patients from the Yale hospital and the Hopkins hospital and whatever sort of large medical centers, at least they might be trying to enroll patients from a broader spectrum across the U.S.

Unfortunately, though, as much as I want to do things
for design often, we’re in a case where there’s a study
that’s already been conducted and we are just sort of stuck analyzing it.
And we wanna get a sense for how representative the results might be for a population.
Sometimes people, when I talk about this, people are like, well, isn’t this what meta-analysis does?
Like meta-analysis enables you to combine multiple randomized trials and come up with sort of an overall effect estimate.
And my answer to that is sort of yes maybe, or no maybe.
Basically, the challenge with meta-analysis, is that until recently, no one really had a potential target population.
It was not very formal about what the target population is.
I think underlying that analysis is generally sort of a belief that the effects are constant and we’re just trying to pool data.
And even just like, you can sort of see this,
like if all of the trials sampled the same non-representative population,
combining them is not going to help you get towards representativeness.
That’s that I have a former Postdoc Hwanhee Hong,
who’s now at Duke. And she has been doing some work to try to bridge these worlds and sort of really try to think through, well, how can we better use multiple trials to get to target population effects? There’s another field it’s called risk cross-design synthesis or research synthesis. This is sort of neat. It’s one where you kind of combine randomized trial data, which might be not representative with non-experimental study data. So sort of explicitly trading off the internal and external validity. I’m not gonna get into the details, there’s some references here. Ellie Kaizar at Ohio State, is one of the people that’s done a lot of work on this. And part of the reason I’m not focused on this is that I work in a lot of areas like education and public health, sort of social science areas, where we often don’t have multiple studies. So we often are stuck with just one study and we’re trying to use that to learn about target populations. So I’m gonna briefly talk about an example where we trying to sort of do this. And basically, the fundamental idea is to re-weight the study sample to look like the target population.
This idea is related to post stratification or, oh my gosh, I’m blanking now.

Raking adjustments in surveys.

So post stratification would be sort of at a simple level,

would be something like...

Well, if we know that males and females have different effects, or let’s say young and old have different effects, let’s estimate the effects separately for young versus old.

And then re-weight those using the population proportions of sort of young versus old.

That sort of stratification doesn’t work if you have more than one or two categorical effect moderators.

And so, what I’m gonna show today is an approach where we use weighting, where we fit a model, predicting participation in the trial, and then weight the trial sample to look like the target population.

So similar idea to things like propensity score weights or non-response adjustment weights in samples.

There is a different approach,

So what I’m gonna illustrate today is sort of this selection weighting strategy.

You also can tackle this external validity by trying to model the outcome very flexibly.
and then project outcomes in the population. In some work I did with Jennifer Hill and others, we showed that BARTs, Bayesian Additive Regression Trees can actually work quite well for that purpose. And more recently, Issa Dahabreh at Brown has done some nice work sort of bridging these two and showing basically a doubly robust kind of idea where we can use both the sample membership model and the outcome model to have better performance. But today, I’m gonna just illustrate the weighting approach, partly because it’s a really nice sort of pedagogical example and helps you kind of see what’s going on in the data. Okay, any questions before I continue? Okay. So the example I’m gonna use is... There was this, I mean, some of you probably know much more about HIV treatment than I do, but the ACTG Trial, which was now quite an old trial, but it was one of the ones that basically showed that HAART therapy, highly active antiretroviral therapy was quite effective at reducing time to AIDS or death compared to standard combination therapy at the time.
So it randomized about 1200 U.S. HIV positive adults to treatment versus control. And the intent to tree analysis in the trial had a hazard ratio of 0.51. So again, very effective at reducing time to AIDS or death. So Steve Cole and I thought kind of asked the question, well, we don’t necessarily just care about the people in the trial. This seems to be a very effective treatment. What could we use this data to project out sort of what the effects of the treatment would be if it were implemented nationwide?

So we from CDC got estimates of the number of people newly infected with HIV in 2006. And basically, asked the question sort of if hypothetically, everyone in that group were able to get HAART versus standard combination therapy, what would be the population impacts of this treatment? In this case, because of sort of data availability, we only had the joint distribution of age, sex and race for the population. So we made sort of a pseudo population, again, sort of representing the U.S. population of newly infected people. But again, all we have is sex, race and age,
So you can see for example, that the trial tended to have more sort of 30 to 39 year olds, many fewer people under 30. The trial had more males and also had more whites and fewer blacks, Hispanic was similar. But I wanna flag and we’ll come back to this in a minute that, in what I’m gonna show, we can adjust for the age, sex, race distribution. But, there’s a real limitation, which is that the CD4 cell count as sort of a measure of disease severity is not available in the population. So this is a potential effect moderator, which we don’t observe in the population. So in sort of projecting the impacts, we can say, well, here is the predicted impact given the age, sex, race distribution, but there’s this unobserved potential effect moderator that we sort of might be worried about kind of in the back of our heads. So again, I briefly mentioned this, this is like the super basic description of what can be done. There are more nuances and I have some sites at the end for sort of more details.
But basically fundamentally will, again, we sort of think about it as we kind of stack our data sets together. So we put our trial sample and our population data set together. We have an indicator for whether someone is in the trial versus the population. And then, we’re gonna wait the trial members by their inverse probability of being in the trial as a function of the observed covariance. And again, very similar intuition and ideas and theory underlying this as underlying things like Horvitz-Thomson estimation in sample surveys and inverse probability of treatment waiting in non-experimental studies. So I showed you earlier that age, sex and race are all related to participation in the trial. are all related to participation in the trial. What I’m not showing you the details of, but just trust me is that those factors also moderate effects in the trial. So the trial showed the largest effects for those ages, 30 to 39, males and black individuals. And so, this is exactly why then what we might think that the overall trial estimate might not reflect what we would see population-wide. Ironically though, it turns out actually it kind of all cancels out.
So this table shows the estimated population effects.

So the first row again, is just the sort of naive trial results. We can then sort of weight by each characteristic separately, and then the bottom row is the combined age, sex, race adjustments.

And you can see sort of actually the hazard ratio was remarkably similar. It’s partly because like the age weightings sort of makes the impact smaller, but then the race weighting makes it bigger. And so then it kind of just washes out.

But again, it’s sort of a nice example, cause you can sort of see how the patterns evolve based on the size of the effects and the sample selection.

I also wanna point out though that, of course, the confidence interval is wider, and that is sort of reflecting the fact that we are doing this extrapolation from the trial sample to the population.

And so there’s sort of a variance price we’ll pay for that.

I haven’t been super formal on the assumptions, but I’m I alluded to this? So I wanna just take a few minutes to turn to what about unobserved moderators?

Because again, we can interpret this 0.57
as the sort of overall population effect estimate
only under an assumption that there are no unobserved
moderators that differ between sample and population,

once we adjust for age, sex, race.

Okay, and in reality, such unobserved effect moderators are likely the rule,

not the exception.

So again, sort of, as I just said, the key assumption is that we’ve basically adjusted
for all of the effect moderators.

Very kind of comparable assumption to the assumption

of no an observed confounding in a non-experimental study.

And one of the reasons this is an important assumption

to think about, is that, it is quite rare actually

to have extensive covariate data overlap

between the sample and the population.

I have been working in this area for...

How many years now?

At least 10 years.

And I’ve found time and time again,

across a number of content areas,

that it is quite rare to have a randomized trial sample

and the target population dataset

with very many comparable measures.

So in the Stuart and Rhodes paper,

this was in like early childhood setting
and each data set, the trial and the population data had like over 400 variables observed at baseline. There were literally only seven that were measured consistently between the two samples. So essentially we have very limited ability then to adjust literally only seven that were measured consistently between the two samples. So essentially we have very limited ability then to adjust for these factors because they just don’t have much overlap. So what that then motivated us to create some sensitivity analysis to basically probe and say, well, what if there is an unobserved effect moderator, how much would that change our population effect estimate? Again, this is very comparable to analysis of sensitivity, to unobserved confounding and non-experimental studies sort of adapted for this purpose of trial population, generalized ability. I think I can skip this in the interest of time and not go through all the details. If anyone wants the slides by the way, feel free to email me, I’m happy to send them. I’m gonna skip this too cause I’ve already said I’m gonna skip this too cause I’ve already said sort of the key assumption that is relevant for right now, but basically what we propose is, I’m gonna talk about two cases. So the easier case is this one where we’re gonna assume that the randomized trial observes all of the effect
And the issue is that our target population dataset does not have some moderators observed. I think this is fairly realistic because at least like to think that the people running the randomized trials have enough scientific knowledge and expertise that they sort of know what the likely effect moderators are and that they measure them in the trial. That is probably not fully realistic, but I'm... I like to give them sort of the benefit of the doubt. And that sort of that's what the ACTG example, where we have CD4 count in the trial, but we just don’t have it in the population. So what we showed is that there’s actually a couple of different ways you can implement this sort of sensitivity analysis. One is essentially kind of an outcome model based one basically, we just sort of specify a range for the unobserved moderator V in the population. So we kind of say, well, we don’t know the distribution of this moderator in the population, but we’re gonna guess that it’s in some range. And then, we kind of projected out using data from the trial to understand like the extent of the moderation.
due to that variable.
There’s another variation on this, which is sort of the weighting variation
where you kind of adjust the weights, essentially again for this unobserved moderator.
Again, either way you sort of basically just have to specify
a potential range for this V, the unobserved moderator
in the population.
So here’s an example of that.
This is a different example, where we were looking at the effects of a smoking cessation intervention
among people in substance use treatment.
And in the randomized trial, the mean addiction score
was four.
But we didn’t have this addiction score, in the target population of interest.
And so, what the sensitivity analysis allows us to do
is to say, well, let’s imagine that range is anywhere from three to five.
And how much does that change our population effect?
Essentially, how steep this line is, is gonna be
sort of determine how much it matters.
And the steepness of the line basically
is how much of a moderator is it,
sort of how much effect heterogeneity is there in the trial
as a result of that variable.
But again, this is at least one way to sort of turn this sort of worry about an unobserved moderator into a more formal statement about how much it really might matter. I'm not gonna get into this partly, so you might also be thinking, well, what if the trial doesn’t know what all the moderators are? And what if there’s some fully unobserved moderator that will call U? This is a much much harder, basically, if anyone wants to try to dig into it, that would be great. Part of the reason it’s harder is because you have to make very strong assumptions about the distribution of the observed covariance and U together. We put out one approach, but it is a fairly special case and not very general. So again, hopefully we’re not in this sort of scenario very often. This is a little bit of a technicality, but often epidemiologists ask this question. So I’ve laid stuff out again with respect to kind of a risk difference or a difference in outcomes and sort of like more of like an additive treatment scale. There is this real complication that arises, which is that if you have like a binary, like the scale of the outcome matters in terms of effect.
And in particular, there might be sort of more apparent effect heterogeneity on one scale versus another. So I'm just kind of flagging this, that like this exists, there are some people sort of looking at this in more formal, but again for now sort of just think about like risk difference kind of scale.

Okay, great. So let me just conclude with a few kind of final thoughts. So, I think all of us, not all of us, but often we sort of want to assume that study results generalize. Often people write a discussion section in a paper, where they kind of qualitatively have some sentences about why they do or don’t think that the results in this paper kind of extend to other groups or other populations. But I think until the past again, sort of five or so years, a lot of that discussion was very hand-wavy and sort of qualitative. I think that what we are seeing in epidemiology and statistics and bias statistics recently has been a push towards having more ability to quantify this and make it sort of more formal statements.
So I think if we wanna be serious though, about assessing and enhancing external validity, again, we really need these different pieces. We need information on the factors that influence heterogeneity the moderators. We need information on the factors that influence participation in rigorous studies like randomized trials. And we need data on all of those things, in the trial and the population. And then finally, we need statistical methods that allow us to use that data to estimate population treatment effects. I would argue that that last bullet is sort of much further along than any of the others. That in my experience, the limiting factor is usually not the methods. The limiting factor at this point in time is the data and sort of the scientific knowledge about these different factors. And that’s what this slide is. So I think I’ve already said, but that again, is sort of one of the motivations for the sensitivity analysis is just a recognition that it’s often, really quite hard to get data that is consistently measured between a trial and a population. So on that point, recommendations again, if we wanna be serious about effect heterogeneity
00:43:51.340 –> 00:43:54.780 or about estimating population treatment effects, we need better information on treatment effect heterogeneity
00:43:59.210 –> 00:44:01.690 that might be better analysis of existing trials, that might be meta-analysis of existing trials.
00:44:04.500 –> 00:44:07.440 That might also be theoretical models for the interventions
00:44:07.440 –> 00:44:10.773 to understand what the likely moderators are.
00:44:11.830 –> 00:44:14.040 We also need better information on the factors
00:44:14.040 –> 00:44:17.160 that influence participation in trials and more discussion
00:44:17.160 –> 00:44:19.913 of how trial samples are selected.
00:44:21.860 –> 00:44:23.330 We need to standardize measures.
00:44:23.330 –> 00:44:26.250 So again, it’s incredibly frustrating when you have trial
00:44:26.250 –> 00:44:29.660 and population data, but the measures in them are not
00:44:29.660 –> 00:44:30.890 consistent.
00:44:30.890 –> 00:44:33.440 There are methods that can be used for this,
00:44:33.440 –> 00:44:35.453 some data harmonization approaches,
00:44:36.390 –> 00:44:38.860 but, they require assumptions.
00:44:38.860 –> 00:44:42.450 It’s better if we can be thoughtful and strategic about,
00:44:42.450 –> 00:44:45.250 for example, common measures across studies.
00:44:45.250 –> 00:44:47.070 I will say one of the frustrations too,
00:44:47.070 –> 00:44:50.830 is that in some fields like the early childhood data
00:44:50.830 –> 00:44:52.070 I talked about,
00:44:52.070 –> 00:44:54.560 part of the problem was like the two data sets might
00:44:54.560 –> 00:44:56.440 actually have the same measure,
00:44:56.440 –> 00:44:58.410 but they didn’t give the raw data,
00:44:58.410 –> 00:45:00.630 and they’re like standardized scales differently.
Like they standardized them to their own population, not sort of more generally. And so they, weren’t sort of on the same scale in the end. As a statistician, of course, I will say we do need more research on the methods and understanding when they work and when they don’t. There are some pretty strong assumptions in these approaches. But again, I think that sort of in some ways, that is further along and then some of the data situations. So I just wanted to take one minute to flag some current work in case partly if anyone wants to ask questions about these.

One thing I’m kind of excited about, especially in my education world is... So what I’ve been talking about today has mostly been, if we have a trial sample and we wanna project to kind of a larger target population. But there’s an equally interesting question, which is sort of how well can randomized trial informs or local decision making? So if we have a randomized trial with 60 schools in it, how well can the results from that trial be used to inform
individual school districts decisions? 

Turns out, not particularly well.

(laughs)

We can talk more about that.

I mentioned earlier, Issa Dahabreh, who’s at Brown,

and he’s really interested in developing sort of the formal

theories underlying different ways of estimating these population effects, again, including some doubly robust approaches.

Trang Nguyen, who works at Hopkins with me,

we are still looking at sort of the sensitivity analysis for unobserved moderators.

I mentioned Hwanhee Hong already, who’s now at Duke.

And she, again, sort of straddles the meta-analysis world

in this world, which has some really interesting connections.

My former student now he’s at Flatiron Health as of a few months ago.

Ben Ackerman, did some work on sort of measurement error and partly how to deal with some of these measurement challenges between the sample and population.

And then I’ll just briefly mention Daniel Westreich at UNC,

who is really...

If you come from sort of more of an epidemiology world,

Daniel has some really nice papers that are sort of trying
to translate these ideas to epidemiology, and this concept of what he calls target validity. So sort of rather than thinking about internal and external validity separately, and as potentially, in kind of conflict with each other, instead really think carefully about a target of inference and then thinking of internal and external validity sort of within that and not sort of trying to prioritize one over the other. And then just an aside, one thing, I would love to do more in the coming years is thinking about combining experimental and non-experimental evidence. I think that is probably where it would be very beneficial to go instead of more of that cross designed synthesis kind of idea. But again, I wanna conclude with this, which is gets us back to design and that again, sort of what is often the limiting factor here is the data and just sort of strong designs. So Rubin, 2005 with better data, fewer assumptions are needed and then Light, Singer and Willett, who are sort of big education methodologists. You can’t fix by analysis what you’ve bungled by design.
So again, just wanna highlight that if we wanna be serious about estimating population effects, we need to be serious about that in our study designs, both in terms of who we recruit, but then also what variables we collect on them. If we do that, I think that we can have the potential to really help guide policy and practice by thinking more carefully about the populations that we care about. Here’s this, there’s my email, if you wanna email me. And thanks to various funders, and then I’ll leave this up for a couple minutes, which are all big, tiny font, some of the references, but then I’ll take that down in a minute so that we can see each other more. So thank you, and I’m very happy to take some questions. I don’t know if you all have a way to organize or people just can jump in. So maybe I’ll ask the question. Thanks Liz, for this very interesting and great talk. So I noticed that you’ve talked about the target population in this framework.
And I think there are situations where the population sample is actually a survey from a larger population. Yeah. Cause we do not really afford to absorb everything, actual population, which will contain like millions of individuals. And so in that situation, does the framework still apply particularly in terms of the sensitivity analysis? And is there any caveat that we should also know in dealing with those data? Great question. And actually, thank you for asking that because I forgot to mention that Ben Ackerman’s dissertation, also looked at that. So I mentioned his measurement error stuff. But yes, actually, so Ben’s second dissertation paper did exactly that, where we sort of laid out the theory for when these the target population data comes from a complex survey itself. Short answer is yes, it all still works. Like you have to use the weights, there are some nuances, but, and you’re right, like essentially, especially like in... Like for representing the U.S. population, often, the data
we have is like the National Health Interview Survey
or the Add Health Survey of Adolescents,
which are these complex surveys.
So short answer is, yeah, it still can work.
Your question about the sensitivity analysis is
actually a really good one and we have not extended...
I'd have to think, I don’t know, off hand, like,
I think it would be sort of straightforward to extend
the sensitivity analysis to that, but we haven’t
actually done it.
Thanks Liz.
The other short question is that I noticed that
in your slide, you first define, PATE as population
but then in one slide you have this Tate,
which I assume is target ate.
And so, I’m just really curious as to like, is there
any, like differences or nuances in the choice of this
terminology?
Good question.
And no, yeah, I’m not...
I wasn’t very precise with that, but in my mind, no.
Over time I’ve been trying to use Tate,
but you can see that kind of just by default,
I still sometimes use PATE.
Part of the reason I use Tate is because I think
the target is just a slightly more general term.
Like people sometimes I think, think if we meet, if we say PATE, the population has to be like the U.S. population or some like very sort of big, very official population in some sense. Whereas, the target average treatment effect, Tate terminology, I think reflects that sometimes it’s just a target group that’s well-defined.

And I think we have a question coming from the chat as well.

So I think, what you mean is sort of, what is the formal theory for projecting the impact to the target population?

That is exactly what some of those people that I referenced

sort of lay out.

Like I didn’t... For this talk, I didn’t get into all the theoretical weeds,

but if you’re interested in that stuff,

probably some of Issa Dahabreh’s work would be the most relevant to look at.
Cause he really lays out sort of the formal theory. I mean, some of my early papers on this topic did it, but his is like a little bit more formal and sort of makes connections to the doubly robust literature and things like that. Anyway, that’s what this whole literature and part of it is sort of building is that theoretical base for doing this. Any other questions?

So in most of the studies that I would work on, they don’t do really have a great idea about what really the population is and how really to measure what really the population is and how really to measure those. So it’s great if I have some measure of the population, but most of the time it is the studies that I work. I have no real measurements on that population. What happens then?

Yeah, great question. And in part, I meant to say this, but that’s one of the reasons why the analogy...
Why the design strategies don’t always work particularly well is like, especially when you’re just starting out a study, right? We don’t really know the target population. I think certainly to do any of these procedures, you need eventually to have a well defined population. But I think that’s partly why some of the analysis approaches are useful is that, you might have multiple target populations. Like we might have one trial, and we might be interested in saying, how well does this generalize to the State of New Hampshire or the State of Vermont or the State of Connecticut? And so, you could imagine one study that’s used to inform multiple target populations. With different assumptions, sort of you have to think through the assumptions for each one. If you don’t even, I guess I would say if you don’t even know who your population is, you shouldn’t be using these methods at all, cause like the whole premise is that there is some well-defined target population and you do need data on it or at least... Yeah, the joint distribution of some covariance...
or something.
Without that, you’re kind of just like,
I don’t know, what a good analogy is,
but you’re kinda just like guessing at everything.
(mumbles)
-No, go ahead.
Go ahead.
- Oh, Vinod, yeah.
All my friends are popping up, it’s great.
(laughs)
- Can I go ahead?
I feel like I’m talking to someone.
- Yeah, go ahead Vinod.
That was a great talk.
So I have a little ill formulated question,
it’s queuing after just the last question
that was asked is,
in clinical set populations where,
in some ways we’re using this clinical samples
to learn about the population because unless they seek help,
we often don’t know what they are in the wild, so to speak.
And so, each sampling of that clinical population
is a maybe by sampling of that larger population
in the wild.
So I guess my question is, how do you get around this,
I guess Rumsfeld problem, which is every time you sample
there’s this unknown, unknown, but there’s no way to get
at them because in some ways, your sampling relies on...

If we could say it relies on help seeking, which is by itself a process. And if we could just stipulate, there's no way to get around that.

How do you see this going forward? Yeah, good question.

I think particularly relevant in mental health research where there's a lot of people who are not seeking treatment. These methods are not gonna help with that in a sense. They are gonna be sort of tuned to whatever population you have.

If you really wanna be thoughtful about that's problem, then we really have to tackle that very creatively and with a lot of resources probably. Surveys, where they would go door to door and knock on doors and do diagnostic interviews. Like if we wanna be really serious about trying to reach everyone and get an estimate of the really sort of true population, then we really have to tackle that very creatively and with a lot of resources probably.
Hi Liz?

Yeah, it’s gonna be a true question and great talk.

by the way.

I’m curious, you mentioned there could be a slight
difference between the terms transportability
and generalizability.

Yeah, I’m curious about that.

- Yeah, briefly, this is a little bit of a…

What’s the word?

Simplification, but briefly I think of generalizability
as one where the sample that, like the trial sample
is a proper subset of the population.

So we do a trial in New Hampshire,
and we’re trying to generalize to new England.

Whereas transportability is one where it is not a
proper
subset, so we do a trial in the United States
and we wanna transport to Europe.

Underlying both, the reason I don’t worry too
much about it,
the terms is because either way,
the assumption is essentially the same.

Like you still have to make this assumption about
no unobserved moderators.

It’s just that it’s probably gonna be a stronger
assumption
and harder to believe,
when transporting rather than when generalizing.

Cause you sort of know that you’re going from one
place
to another in some sense.
00:58:39.380 –> 00:58:40.500 - Thanks, makes sense.
00:58:40.500 –> 00:58:41.333 - Sure.
00:58:42.560 –> 00:58:44.540 - I think there’s another question in the chat.
00:58:44.540 –> 00:58:46.410 - Yeah, so this is a great question.
00:58:46.410 –> 00:58:48.400 I’m glad shows you on.
00:58:48.400 –> 00:58:50.220 I hope I got that.
00:58:50.220 –> 00:58:52.530 It seems there are multiple ways to calculate the Tate
00:58:52.530 –> 00:58:55.420 from standardization to waiting to the outcome model.
00:58:55.420 –> 00:58:57.420 Do you have comments for their performance under different
00:58:57.420 –> 00:58:58.420 circumstances?
00:58:58.420 –> 00:59:00.590 Great question, and I don’t.
00:59:00.590 –> 00:59:01.890 I mean, there has been...
00:59:01.890 –> 00:59:03.900 This is an area where I think
00:59:03.900 –> 00:59:06.300 it’d be great to have more research on this topic.
00:59:06.300 –> 00:59:09.490 So I have this one paper with Holger Kern and Jennifer Hill
00:59:09.490 –> 00:59:14.080 where we sort of did try to kind of explore that.
00:59:14.080 –> 00:59:16.090 And honestly, what we found not surprisingly
00:59:16.090 –> 00:59:20.080 is that if that no unmeasured moderator assumption holds,
00:59:20.080 –> 00:59:22.650 all the different methods are pretty good and fine.
00:59:22.650 –> 00:59:25.030 And like, we didn’t see much difference in them.
00:59:25.030 –> 00:59:27.650 If that no unobserved moderator assumption doesn’t hold
00:59:27.650 –> 00:59:28.840 then of course, none of them are good.
00:59:28.840 –> 00:59:31.843 So it sort of is like similar to propensity score world.
00:59:33.097 –> 00:59:35.240 Like, the data you have is more important than what you do
00:59:35.240 –> 00:59:36.653 with the data in a sense.
But anyway, I think that is something that like,

we need a lot more work on.

One thing, for example, I do have a student working on this.

Like, we’re trying to see if your sample

is a tiny proportion of the population, like how...

Cause like there’s different.

That’s one where like waiting might not work as well

actually, who knows.

Anyways, so like all of these different data scenarios,

I think need a lot more investigation to have better

guidance on when the different methods work well.

Anything else or maybe we’re out of time?

I don’t know, how tight you are at one o’clock.

- I think we’re at an hour, so let’s...