0:00:00 –> 0:00:02.4 - Maybe one or two minutes and then,

0:00:02.4 –> 0:00:03.233 I'll have you introduced.

0:00:03.233 –> 0:00:04.64 - And it's about, and so I. . .

0:00:04.64 –> 0:00:06.86 And it's gonna be more fun for me if it's a little

0:00:06.86 –> 0:00:08.51 interactive, as much as we can make it.

0:00:08.51 –> 0:00:11.76 So I won't be able to see all of you nodding and whatnot,

0:00:11.76 –> 0:00:14.827 but please feel free to jump in.

0:00:14.827 –> 0:00:16.83 And the talk's gonna be pretty non-technical.

0:00:16.83 –> 0:00:18.96 My goal is mostly to sort of help

0:00:18.96 –> 0:00:23.36 convey some of the concepts and ideas and so I will.

0:00:23.36 –> 0:00:27.143 Hopefully it will be a reasonable topic to do via Zoom.

0:00:30.05 –> 0:00:31.42 Great, so I think,

0:00:32.65 –> 0:00:35.67 Frank basically gave this stuff that's relevant

0:00:35.67 –> 0:00:36.8 on this slide.

0:00:36.8 –> 0:00:38.89 I do also wanna apologize, those of you guys

0:00:38.89 –> 0:00:41.17 who I was supposed to meet with this morning, we have a. . .

0:00:41.17 –> 0:00:43.85 My husband broke his collarbone over the weekend.

0:00:43.85 –> 0:00:46.84 So I've had to cancel things this morning,

0:00:46.84 –> 0:00:49.853 but I'm glad I'm able to still do this seminar,

0:00:51.278 –> 0:00:52.38 I didn't wanna,

0:00:52.38 –> 0:00:53.38 have to cancel that.

0:00:54.35 –> 0:00:56.06 So again,

0:00:56.06 –> 0:00:58.53 the topic is gonna be sort of this idea of external

0:00:58.53 –> 0:01:01.33 validity, which I think is a topic that people often

0:01:01.33 –> 0:01:03.54 are interested in because it's the sort of thing

0:01:03.54 –> 0:01:06.17 that we often think sort of qualitatively about,

0:01:06.17 –> 0:01:08.31 but there hasn't been a lot of work thinking about it

0:01:08.31 –> 0:01:09.143 quantitatively.

0:01:09.143 –> 0:01:11.38 So again, my goal today will be to sort of help

0:01:11.38 –> 0:01:14.84 give a framework for thinking about external validity

0:01:14.84 –> 0:01:16.863 in sort of a more formal way.

0:01:18.9 –> 0:01:22.38 So let's start out with the sorts of questions

0:01:22.38 –> 0:01:25.22 that might be relevant when you're thinking about

0:01:25.22 –> 0:01:26.74 external validity.

0:01:26.74 –> 0:01:30.08 So it might be research questions like a health insurer

0:01:30.08 –> 0:01:33.72 is deciding whether or not to approve some new treatment

0:01:33.72 –> 0:01:35.89 for back pain.

0:01:35.89 –> 0:01:39.09 There might be interested predicting overall population

0:01:39.09 –> 0:01:43.14 impacts of a broad public health media campaign.

0:01:43.14 –> 0:01:46.13 A physician practice might be deciding whether training

0:01:46.13 –> 0:01:48.77 providers in a new intervention would actually be cost

0:01:48.77 –> 0:01:52.63 effective given the patient population that they have.

0:01:52.63 –> 0:01:54.97 And that I felt like I needed to get some COVID

0:01:54.97 –> 0:01:57.3 example in...

0:01:57.3 –> 0:01:59.29 But, for example, a healthcare system,

0:01:59.29 –> 0:02:02.08 might wanna know whether it's sort of giving convalescent

0:02:02.08 –> 0:02:05.69 plasma to all of the individuals recently diagnosed

0:02:05.69 –> 0:02:08.24 with COVID-19 in their system, whether that would

0:02:08.24 –> 0:02:10.853 sort of lead to better outcomes overall.

0:02:12.47 –> 0:02:14.56 So all of these...

0:02:14.56 –> 0:02:16.86 What I'm distinguishing here or sort of trying to convey

0:02:16.86 –> 0:02:19.88 is that all of these reflect what I will call a population

0:02:19.88 –> 0:02:21.5 average treatment effect.

0:02:21.5 –> 0:02:24.64 So across some well-defined population,

0:02:24.64 –> 0:02:28.24 does some intervention work sort of on average.

0:02:28.24 –> 0:02:30.21 The population might be pretty narrow.

0:02:30.21 –> 0:02:33.13 Again, it might be the patients in one particular

0:02:33.13 –> 0:02:35.49 physician practice, or might be quite broad.

0:02:35.49 –> 0:02:38.14 It could be everyone in the State of Connecticut

0:02:38.14 –> 0:02:40.39 or in the entire country.

0:02:40.39 –> 0:02:44.24 But either way, it's a well-defined kind of population
0:02:44.24 –> 0:02:46.08 and we'll come back to that.
0:02:46.08 –> 0:02:47.5 What's really important,
0:02:47.5 –> 0:02:50.02 and this will sort of underlie much of the talk
0:02:50.02 –> 0:02:52.48 is that kind of the whole point is that there might
0:02:52.48 –> 0:02:54.61 be underlying treatment effect heterogeneity.
0:02:54.61 –> 0:02:56.89 So there might be some individuals
0:02:56.89 –> 0:02:59.1 for whom this treatment of interest is actually
0:02:59.1 –> 0:03:01.07 more effective than others.
0:03:01.07 –> 0:03:04.41 But what I wanna be clear about, is the goal of inference
0:03:04.41 –> 0:03:06.98 that I'm talking about today, is gonna be about
0:03:06.98 –> 0:03:08.75 this overall population average.
0:03:08.75 –> 0:03:11.45 So we're not trying to say like which people
0:03:11.45 –> 0:03:14.41 are gonna benefit more or sort of to which people
0:03:14.41 –> 0:03:15.97 should we give this treatment.
0:03:15.97 –> 0:03:19.56 It's really more a question of sort of more population
0:03:19.56 –> 0:03:21.53 level decisions, sort of if we have. . .
0:03:21.53 –> 0:03:23.65 If we're making a decision, that's sort of a policy
0:03:23.65 –> 0:03:25.25 kind of population level,
0:03:25.25 –> 0:03:28.35 on average is this gonna be something that makes sense.
0:03:28.35 –> 0:03:30.42 So I hope that distinction makes sense.
0:03:30.42 –> 0:03:32.343 I'm happy to come back to that.
0:03:35.36 –> 0:03:38.243 So again until I don't know, five or,
0:03:38.243 –> 0:03:41.09 well maybe now more than 10 years ago,
0:03:41.09 –> 0:03:42.99 there had been relatively little attention
0:03:42.99 –> 0:03:46.47 to the question of how well results from
0:03:46.47 –> 0:03:50.04 kind of well-designed studies like a randomized trial
0:03:50.04 –> 0:03:52.92 might carry over to a relevant target population.
0:03:52.92 –> 0:03:55.83 I think in much of statistics as well as fields
0:03:55.83 –> 0:04:00.12 like education research, public policy, even healthcare,
0:04:00.12 –> 0:04:02.56 there's really been a focus on randomized trials

0:04:02.56 –> 0:04:04.95 and getting internal validity,

0:04:04.95 –> 0:04:07.44 and I'll formalize this in a minute.

0:04:07.44 –> 0:04:09.93 But in the past 10 or so years, there's been more and more

0:04:09.93 –> 0:04:13.18 interest in this idea of how well can we take the results

0:04:13.18 –> 0:04:17.03 from a particular study and then project them

0:04:17.03 –> 0:04:19.62 to well-defined target population.

0:04:19.62 –> 0:04:21.33 And again, so today I'm gonna try to give

0:04:21.33 –> 0:04:24.1 sort of an overview of the thinking in this area,

0:04:24.1 –> 0:04:26.93 along with some of the limitations and in particular,

0:04:26.93 –> 0:04:29.78 the data limitations that we have in thinking about this.

0:04:32.84 –> 0:04:35.72 One thing I do wanna be clear about is there's a lot

0:04:35.72 –> 0:04:38.01 of reasons why results from randomized trials

0:04:38.01 –> 0:04:39.58 might not generalize.

0:04:39.58 –> 0:04:42.32 There's some classic examples in education

0:04:42.32 –> 0:04:44.45 where there are scale-up problems.

0:04:44.45 –> 0:04:47.903 The classic example is one I'm looking at,

0:04:49.89 –> 0:04:50.75 class size.

0:04:50.75 –> 0:04:53.88 And so, in Tennessee, they randomly assign kids

0:04:53.88 –> 0:04:56.62 to be in smaller versus larger classes

0:04:56.62 –> 0:04:59.57 and found quite large effects of smaller classes.

0:04:59.57 –> 0:05:02.53 But then, when the State of California tried to implement

0:05:02.53 –> 0:05:05.88 this, the problem is that you need a lot more teachers

0:05:05.88 –> 0:05:08.04 to kind of roll that out statewide.

0:05:08.04 –> 0:05:10.72 And so, it led actually to a different pool of teachers

0:05:10.72 –> 0:05:11.553 being hired.

0:05:11.553 –> 0:05:13.97 And so, there's sort of scale-up problems

0:05:13.97 –> 0:05:16.17 sometimes with the interventions and that might lead

0:05:16.17 –> 0:05:19.01 to different contexts or different implementation.

0:05:19.01 –> 0:05:21.25 Today, what I'm gonna be focusing on are differences

0:05:21.25 –> 0:05:23.503 between a sample and a population.

0:05:24.77 –> 0:05:27.63 Their difference is in sort of baseline characteristics,

0:05:27.63 –> 0:05:28.757 that moderate treatment effects.

0:05:28.757 –> 0:05:31.763 And again, I'll formalize this a little bit as we go along.

0:05:32.83 –> 0:05:34.23 Just as a little bit of an aside,

0:05:34.23 –> 0:05:36.83 but in case some of you know this field a little bit,

0:05:36.83 –> 0:05:38.74 just to give you a little, just. . .

0:05:38.74 –> 0:05:40 I wanna flag this.

0:05:40 –> 0:05:42.81 Some people might use the term transportability.

0:05:42.81 –> 0:05:45.72 So some of the literature in this field uses the term

0:05:45.72 –> 0:05:47.17 transportability.

0:05:47.17 –> 0:05:50.09 I tend to use generalizability.

0:05:50.09 –> 0:05:51.92 There's some subtle differences between the two,

0:05:51.92 –> 0:05:55.46 which we can come back to, but for all intents and purposes,

0:05:55.46 –> 0:05:58.66 like they basically can think of them interchangeably

0:05:58.66 –> 0:06:00.21 for now.

0:06:00.21 –> 0:06:02.05 I also wanna note, if any of you kind of come

0:06:02.05 –> 0:06:05.93 from like a survey world, these debates about

0:06:05.93 –> 0:06:09.33 kind of how well a particular sample reflects a target

0:06:09.33 –> 0:06:12.2 population are exactly, not exactly the same,

0:06:12.2 –> 0:06:14.95 but very similar to the debates happening in the survey

0:06:14.95 –> 0:06:18.85 world around non-probability samples and sort of concerns

0:06:18.85 –> 0:06:19.683 about,

0:06:20.85 –> 0:06:24.76 the use of like say online surveys and things that might not

0:06:24.76 –> 0:06:28.35 have a true formal sort of survey sampling design,

0:06:28.35 –> 0:06:30.81 and sort of some of the concerns that arise about

0:06:30.81 –> 0:06:31.643 generalizability.

0:06:31.643 –> 0:06:34.11 So there's this whole parallel literature in the survey

0:06:34.11 –> 0:06:34.99 world.

0:06:34.99 –> 0:06:36.95 Andrew Mercer has a nice summary of that.

0:06:36.95 –> 0:06:39.123 Again, I'm happy to talk more about that.

0:06:41.39 –> 0:06:43.803 Okay, any questions before I keep going?

0:06:48.5 –> 0:06:49.44 Okay.

0:06:49.44 –> 0:06:52.35 So let me formalize kind of what we're talking about

0:06:52.35 –> 0:06:53.48 a little bit.

0:06:53.48 –> 0:06:54.66 This is...

0:06:54.66 –> 0:06:59.2 This framework is now, 12 years old.

0:06:59.2 –> 0:07:00.55 Time goes quickly.

0:07:00.55 –> 0:07:04.66 But we're just to formalize what we're interested in.

0:07:04.66 –> 0:07:07.09 The goal is to estimate, again, this what I'll call

0:07:07.09 –> 0:07:09.483 a population average treatment effect or PATE.

0:07:10.44 –> 0:07:12 And so here,

0:07:12 –> 0:07:14.36 hopefully you're familiar with sort of potential outcomes

0:07:14.36 –> 0:07:15.91 and causal inference.

0:07:15.91 –> 0:07:18.78 But the idea is that we have some well-defined population

0:07:18.78 –> 0:07:20.1 of size N.

0:07:20.1 –> 0:07:23.76 And Y(1) is the potential outcomes, if people

0:07:23.76 –> 0:07:27.79 in that population receive the treatment condition

0:07:27.79 –> 0:07:29.05 of interest.

0:07:29.05 –> 0:07:31.86 Y(0) are the outcomes if they receive the control

0:07:31.86 –> 0:07:33.89 or comparison condition of interest.

0:07:33.89 –> 0:07:35.4 So here, we're just saying we're interested

0:07:35.4 –> 0:07:39.75 in the average effect, basically sort of the difference

0:07:39.75 –> 0:07:44.463 in potential outcomes, average across the population.

0:07:45.53 –> 0:07:49.33 We could be doing this with risk ratios

0:07:49.33 –> 0:07:51.45 or odds ratios or something.

0:07:51.45 –> 0:07:53.15 Those are a little more complicated because the math

0:07:53.15 –> 0:07:55.12 doesn't work as nicely.

0:07:55.12 –> 0:07:57.38 So for now think about it more like risk differences

0:07:57.38 –> 0:07:59.5 or something, if you have a binary outcome,

0:07:59.5 –> 0:08:01.573 the same fundamental points hold.

0:08:02.57 –> 0:08:05.07 So I'm not gonna tell you right now where

0:08:05.07 –> 0:08:08.01 the data we have came from, but imagine that we just

0:08:08.01 –> 0:08:10.51 have a simple estimate of this PATE,

0:08:10.51 –> 0:08:13.67 as the difference in means of some outcome

0:08:13.67 –> 0:08:16.18 between an observed treated group and an observed

0:08:16.18 –> 0:08:17.18 control group.

0:08:17.18 –> 0:08:19.52 So again, we see that there's a bunch of people

0:08:19.52 –> 0:08:22.01 who got treated, a bunch of people who got control,

0:08:22.01 –> 0:08:25.35 and we might estimate this PATE as just the simple

0:08:25.35 –> 0:08:27.85 difference in means between again, the treatment group

0:08:27.85 –> 0:08:29.35 and the control group.

0:08:29.35 –> 0:08:31.56 So what I wanna talk through for the next couple of minutes,

0:08:31.56 –> 0:08:35.93 is the bias in this sort of naive estimate of the PATE.

0:08:35.93 –> 0:08:37.94 So we'll call that Delta.

0:08:37.94 –> 0:08:40.15 So I'm being a little loose with notation here,

0:08:40.15 –> 0:08:43.27 but sort of the PATE that the bias essentially

0:08:43.27 –> 0:08:45.17 think of it as sort of the difference between

0:08:45.17 –> 0:08:49.24 the true population effect and our naive estimate of it.

0:08:49.24 –> 0:08:53.95 And what this paper did with Gary King and Kosuke Imai,

0:08:53.95 –> 0:08:58.38 we sort of laid how different choices of study designs

0:08:58.38 –> 0:09:00.84 impact the size of this bias.

0:09:00.84 –> 0:09:02.61 And in particular, we showed that sort of under

0:09:02.61 –> 0:09:05.47 some simplifying situations,

0:09:05.47 –> 0:09:07.4 sort of mathematical simplicity,

0:09:07.4 –> 0:09:11.08 you can decompose that overall bias into four pieces.

0:09:11.08 –> 0:09:15.36 So the two Delta S terms are what are called,

0:09:15.36 –> 0:09:17.45 what we call sample selection bias.

0:09:17.45 –> 0:09:22.09 So basically, the bias that comes in if our data sample

0:09:22.09 –> 0:09:24.79 is not representative of the target population

0:09:24.79 –> 0:09:25.74 that we care about.

0:09:26.75 –> 0:09:31.3 The Delta T terms are our typical sort of confounding bias.

0:09:31.3 –> 0:09:35.67 So bias that comes in if our treatment group is dissimilar

0:09:35.67 –> 0:09:36.863 from our control group.

0:09:37.87 –> 0:09:40.34 The X refers to the variables we observe,

0:09:40.34 –> 0:09:43.373 and the U refers to variables that we don't observe.

0:09:44.67 –> 0:09:46.28 So what we then did in the paper,

0:09:46.28 –> 0:09:49.22 and this is sort of what motivates a lot of this work

0:09:49.22 –> 0:09:51.37 is to think through these, again, the trade offs

0:09:51.37 –> 0:09:53.2 in these different designs.

0:09:53.2 –> 0:09:56.08 And essentially what we're trying to sort of point out

0:09:56.08 –> 0:09:57.16 is that...

0:09:58.86 –> 0:10:01.19 Let's go to the second row of this table first actually,

0:10:01.19 –> 0:10:02.46 a typical experiment.

0:10:02.46 –> 0:10:05.6 So a typical experiment, I would say is one where

0:10:05.6 –> 0:10:08.05 we kind of take whoever comes in the door,

0:10:08.05 –> 0:10:11.22 we kind of try to recruit people for a randomized trial,

0:10:11.22 –> 0:10:16.22 whether that's schools or patients or whatever it is.

0:10:16.42 –> 0:10:18.81 And we randomized them to treatment and control groups.

0:10:18.81 –> 0:10:21.06 So that is our typical randomized experiment.

0:10:22.1 –> 0:10:26.38 The treatment selection bias in that case is zero.

0:10:26.38 –> 0:10:29.14 In expectation, that's why we like randomized experiments.

0:10:29.14 –> 0:10:31.81 In expectation, there is no confounding

0:10:31.81 –> 0:10:34.3 and we get an unbiased treatment effect estimate

0:10:34.3 –> 0:10:36.67 for the sample at hand.

0:10:36.67 –> 0:10:39.83 The problem for population inference

0:10:39.83 –> 0:10:43.3 is that the Delta S terms might be big,

0:10:43.3 –> 0:10:46.23 because the people that agree to be in a randomized trial,

0:10:46.23 –> 0:10:49.1 might be quite different from the overall population

0:10:49.1 –> 0:10:50.63 that we care about.

0:10:50.63 –> 0:10:53.01 So in this paper, we're trying to just sort of. . .

0:10:53.01 –> 0:10:55.65 In some ways, be a little provocative and point this out

0:10:55.65 –> 0:10:59.43 that our standard thinking about study designs

0:10:59.43 –> 0:11:03.24 and sort of our prioritization of randomized trials,

0:11:03.24 –> 0:11:07.13 implicitly prioritizes internal validity over external

0:11:07.13 –> 0:11:08.4 validity.

0:11:08.4 –> 0:11:12.03 And in particular, if we really care about

0:11:12.03 –> 0:11:15.01 population effects, we really should be thinking about

0:11:15.01 –> 0:11:18.2 these together and trying to sort of have small

0:11:18.2 –> 0:11:21.82 sample selection bias and small treatment selection bias.

0:11:21.82 –> 0:11:25.45 So an ideal experiment would be one where we can randomly

0:11:25.45 –> 0:11:27.61 select people for our trial.

0:11:27.61 –> 0:11:29.84 Let's say we have. . .

0:11:29.84 –> 0:11:31.06 Well, actually, I'll come back to that in a second.

0:11:31.06 –> 0:11:34.02 Randomly select people for our trial and then randomly

0:11:34.02 –> 0:11:36.56 assign people to treatment or control groups.

0:11:36.56 –> 0:11:40.68 And in expectation, we will have zero bias in our population

0:11:40.68 –> 0:11:42.24 effect estimate.

0:11:42.24 –> 0:11:43.97 But these other designs, and again,

0:11:43.97 –> 0:11:47.04 like a typical experiment might end up having larger bias

0:11:47.04 –> 0:11:50.91 overall, than a well designed non-experimental study,

0:11:50.91 –> 0:11:53.65 where if we do a really good job like adjusting

0:11:53.65 –> 0:11:55.25 for confounders,

0:11:55.25 –> 0:11:59.27 it may be that well done non-experimental study

0:11:59.27 –> 0:12:01.94 conducted using say the electronic health records

0:12:01.94 –> 0:12:05.7 from a healthcare system might actually give us lower bias

0:12:05.7 –> 0:12:08.29 for a population effect estimate.

0:12:08.29 –> 0:12:12.12 Then does a non-representative small randomized trial.

0:12:12.12 –> 0:12:13.48 Again, a little provocative,

0:12:13.48 –> 0:12:16.67 but I think useful to be thinking about what is really our

0:12:16.67 –> 0:12:19.34 target of inference and how do we get data that is most

0:12:19.34 –> 0:12:20.513 relevant for that.

0:12:21.57 –> 0:12:24.26 I will also just as a small aside,

0:12:24.26 –> 0:12:25.74 maybe a little on the personal side,

0:12:25.74 –> 0:12:28.43 but it's been striking to me in the past two days.

0:12:28.43 –> 0:12:31.3 So my husband broke his collarbone over the weekend.

0:12:31.3 –> 0:12:34.73 And it turns out the break is one where there's a little bit

0:12:34.73 –> 0:12:37.76 of debate about whether you should have surgery or not.

0:12:37.76 –> 0:12:39.36 Although kind of recent thinking is that

0:12:39.36 –> 0:12:40.29 there should be surgery.

0:12:40.29 –> 0:12:44.24 And I was doing a PubMed search as a good statistician

0:12:44.24 –> 0:12:46.97 public health person whose family member

0:12:46.97 –> 0:12:49.3 needs medical treatment.

0:12:49.3 –> 0:12:51.79 And I found all these randomized trials that actually

0:12:51.79 –> 0:12:54.91 randomized people to get surgery or not.

0:12:54.91 –> 0:12:56 And then I came home. . .

0:12:56 –> 0:12:58.75 Oh, no, I didn't come home, we were home all the time.

0:12:58.75 –> 0:13:00.32 I asked my husband later, I was like,

0:13:00.32 –> 0:13:02.38 would you ever agree to be randomized?

0:13:02.38 –> 0:13:04.72 Like right now, we are trying to make this decision about,

0:13:04.72 –> 0:13:06.77 should you have surgery or not.

0:13:06.77 –> 0:13:09.05 And would we ever agree to be randomized?

0:13:09.05 –> 0:13:11.065 And he's like, no, we wouldn't.

0:13:11.065 –> 0:13:14.55 We're gonna go with what the physician recommends

0:13:14.55 –> 0:13:16.3 and what we feel is comfortable.

0:13:16.3 –> 0:13:19.25 And it really just hit home for me at this point that

0:13:19.25 –> 0:13:22.07 the people who agree to be randomized or the context

0:13:22.07 –> 0:13:25.86 under which we can sort of randomize

0:13:25.86 –> 0:13:27.73 are sometimes fairly limited.

0:13:27.73 –> 0:13:31.23 And again, so partly what this body of research is trying

0:13:31.23 –> 0:13:33.41 to do is sort of think through what are the implications

0:13:33.41 –> 0:13:36.893 of that when we do wanna make population inferences.

0:13:38.23 –> 0:13:39.063 Make sense so far?

0:13:39.063 –> 0:13:41.253 I can't see faces, so hopefully.

0:13:43.29 –> 0:13:44.123 Okay.

0:13:46.5 –> 0:13:47.58 So,

0:13:47.58 –> 0:13:50.27 I will say a lot of my work in this area has actually,

0:13:50.27 –> 0:13:53.48 in part been just helping or trying to raise awareness

0:13:53.48 –> 0:13:55.98 of thinking about external validity bias.

0:13:55.98 –> 0:13:59.9 So some of the research in this area has been trying

0:13:59.9 –> 0:14:02.52 to understand how big of a problem is this.

0:14:02.52 –> 0:14:05.96 If maybe people don't agree to be in randomized trials

0:14:05.96 –> 0:14:07.17 very often,

0:14:07.17 –> 0:14:09.81 but maybe that doesn't really cause bias in terms

0:14:09.81 –> 0:14:12.3 of our population effect estimates.

0:14:12.3 –> 0:14:14.67 So what I've done in a couple of papers on these

0:14:14.67 –> 0:14:18.24 other sides on this slide is basically trying to formalize

0:14:18.24 –> 0:14:22.17 this and it's pretty intuitive, but basically we show,

0:14:22.17 –> 0:14:24.15 and I'm not showing you the formulas here.

0:14:24.15 –> 0:14:27.91 But intuitively, there will be bias in a population effect

0:14:27.91 –> 0:14:31.55 estimate essentially if participation in the trial

0:14:32.59 –> 0:14:35.21 is associated with the size of the impacts.

0:14:35.21 –> 0:14:36.563 So in particular,

0:14:37.51 –> 0:14:39.25 what I'll call the external validity bias.

0:14:39.25 –> 0:14:40.083 So,

0:14:40.083 –> 0:14:42.15 those Delta S terms kind of the bias

0:14:42.15 –> 0:14:44.72 due to the lack of representativeness

0:14:44.72 –> 0:14:47.52 is a function of the variation of the probabilities

0:14:47.52 –> 0:14:49.64 of participating in a trial,

0:14:49.64 –> 0:14:51.54 variation and treatment effects,

0:14:51.54 –> 0:14:54.19 and then the correlation between those things.

0:14:54.19 –> 0:14:55.77 So if constant...

0:14:55.77 –> 0:14:57.64 If we have treat constant treatment effects

0:14:57.64 –> 0:14:59.43 or the treatment effect is zero

0:14:59.43 –> 0:15:02.34 or is two for everyone, there's gonna be no external

0:15:02.34 –> 0:15:03.173 validity bias.

0:15:03.173 –> 0:15:04.96 It doesn't matter who is in our study.

0:15:06.3 –> 0:15:07.52 Or if there...

0:15:07.52 –> 0:15:10.03 If everyone has an equal probability of participating

0:15:10.03 –> 0:15:13.77 in the study, we really do have a nice random selection,

0:15:13.77 –> 0:15:17.12 then again, there's gonna be no external validity bias.

0:15:17.12 –> 0:15:19.89 Or if the factors that influence whether or not you

0:15:19.89 –> 0:15:23.44 participate in the study are independent of the factors

0:15:23.44 –> 0:15:25.15 that moderate treatment effects,

0:15:25.15 –> 0:15:27.803 again, there'll be no external validity bias.

0:15:28.81 –> 0:15:32.25 The problem is that we often have very limited information

0:15:32.25 –> 0:15:33.92 about these pieces.

0:15:33.92 –> 0:15:37.94 We, as a field, I think medicine, public health, education,

0:15:37.94 –> 0:15:41.01 all the fields I worked in, there has not been much

0:15:41.01 –> 0:15:44.2 attention paid to these processes of how we actually

0:15:44.2 –> 0:15:45.97 enroll people in studies.

0:15:45.97 –> 0:15:49.08 And so it's hard to know kind of what factors relate

0:15:49.08 –> 0:15:52.03 to those and if those then also moderate treatment effects.

0:15:53.064 –> 0:15:54.103 (phone ringing)

0:15:54.103 –> 0:15:55.36 Oops, sorry.

0:15:55.36 –> 0:15:57.8 Incoming phone call, which I will ignore.

0:15:57.8 –> 0:15:58.89 So,

0:15:58.89 –> 0:16:00.1 there has been...

0:16:01.01 –> 0:16:01.843 Sorry.

0:16:02.95 –> 0:16:05.31 There has been a little bit of work trying to document this

0:16:05.31 –> 0:16:10.31 in real data and find empirical evidence on these sizes.

0:16:10.78 –> 0:16:13 The problem, and sorry, some of the...

0:16:13 –> 0:16:13.95 Some of you might...

0:16:13.95 –> 0:16:15.82 If any of you are familiar with the, like,

0:16:15.82 –> 0:16:18.23 within what it's called the within study comparison

0:16:18.23 –> 0:16:19.063 literature.

0:16:19.063 –> 0:16:21.75 So there's this whole literature on non-experimental studies

0:16:23.24 –> 0:16:27.57 that sort of try to estimate the bias due to non-random

0:16:27.57 –> 0:16:29.7 treatment assignment.

0:16:29.7 –> 0:16:31.51 This is sort of analogous to that.

0:16:31.51 –> 0:16:33.71 But the problem here is that what you need is you need

0:16:33.71 –> 0:16:37.24 an accurate estimate of the impact in the population.

0:16:37.24 –> 0:16:40.14 And then you also need sort of estimates of the impact

0:16:40.14 –> 0:16:43.69 in samples that are sort of obtained in kind of typical

0:16:43.69 –> 0:16:44.99 ways.

0:16:44.99 –> 0:16:46.69 So that's actually really hard to do.

0:16:46.69 –> 0:16:49.05 So I'll just briefly talk through two examples.

0:16:49.05 –> 0:16:51.81 And if any of you have data examples that you think might

0:16:51.81 –> 0:16:54.57 sort of be useful for generating evidence,

0:16:54.57 –> 0:16:56.8 that would be incredibly useful.

0:16:56.8 –> 0:16:58.88 So one of the examples is...

0:17:00.05 –> 0:17:01.75 So let me back up for a second.

0:17:01.75 –> 0:17:03.33 In the field of mental health research,

0:17:03.33 –> 0:17:05.53 there's been a push recently, or actually not so much

0:17:05.53 –> 0:17:08.27 recently in the past, like 10, 15 years

0:17:08.27 –> 0:17:11.81 to do what I call or what are called pragmatic trials

0:17:11.81 –> 0:17:14.76 with the idea of enrolling much more. . .

0:17:15.91 –> 0:17:20.71 A much broader set of people use a broader set of practices

0:17:20.71 –> 0:17:22.393 or locations around the country.

0:17:23.4 –> 0:17:26.62 And so what this Wisniewski et al people did was they took

0:17:26.62 –> 0:17:28.94 the data from one of those large pragmatic trials.

0:17:28.94 –> 0:17:29.773 And the idea they. . .

0:17:29.773 –> 0:17:32.53 Again, the idea was that it should be more representative

0:17:32.53 –> 0:17:35.07 of people in this case with depression

0:17:35.07 –> 0:17:36.83 across the U.S.

0:17:36.83 –> 0:17:38.1 And then, they said, well, what if. . .

0:17:38.1 –> 0:17:39.56 In fact, we didn't have that.

0:17:39.56 –> 0:17:43.76 What if we use sort of our normal study inclusion

0:17:43.76 –> 0:17:47.36 and exclusion criteria, it's sort of been, we'd like subset,

0:17:47.36 –> 0:17:49.96 this pragmatic trial data to the people that we think

0:17:49.96 –> 0:17:53.26 would have been more typically included in a sort of more

0:17:53.26 –> 0:17:55.22 standard randomized trial.

0:17:55.22 –> 0:17:57.74 And sort of not surprisingly, they found that

0:17:57.74 –> 0:17:59.24 the people in the sort of what they call

0:17:59.24 –> 0:18:02.93 the efficacy sample, those sort of typical trial sample

0:18:02.93 –> 0:18:05.49 had better outcomes and larger treatment effects

0:18:05.49 –> 0:18:08.853 than the overall pragmatic trial sample as a whole.

0:18:10.34 –> 0:18:14.59 We did something similar sort of in education research where

0:18:15.45 –> 0:18:16.48 it's a little bit in the weeds.

0:18:16.48 –> 0:18:17.85 I don't really wanna get into the details,

0:18:17.85 –> 0:18:22.05 but we essentially had a pretty reasonable regression

0:18:22.05 –> 0:18:23.29 discontinuity design.

0:18:23.29 –> 0:18:26.18 So we were able to get estimates of the effects of this

0:18:26.18 –> 0:18:30.03 reading first intervention across a number of states.

0:18:30.03 –> 0:18:33.78 And we then compared those state wide impact estimates

0:18:33.78 –> 0:18:37.69 to the estimates you would get if we enrolled only

0:18:37.69 –> 0:18:40.73 the sorts of schools and school districts that are typically

0:18:40.73 –> 0:18:44.11 included in educational evaluations.

0:18:44.11 –> 0:18:47.64 And there we found that this external validity bias

0:18:47.64 –> 0:18:50.04 was about 0.1 standard deviations,

0:18:50.04 –> 0:18:52.97 which in education world is fairly large.

0:18:52.97 –> 0:18:55.66 Certainly people would be concerned about an internal

0:18:55.66 –> 0:18:57.53 validity bias of that size.

0:18:57.53 –> 0:18:59.71 So we were able to sort of use this to say, look,

0:18:59.71 –> 0:19:03.01 if we really wanna be serious about external validity,

0:19:03.01 –> 0:19:06.4 it might be as much of a problem as sort of typical internal

0:19:06.4 –> 0:19:09.353 validity bias that people care about in that field.

0:19:12.74 –> 0:19:14.53 So again, the problem though, is we don't usually

0:19:14.53 –> 0:19:16.9 have these sorts of designs where we have a population

0:19:16.9 –> 0:19:18.99 effect estimate, and then sample estimates,

0:19:18.99 –> 0:19:20.62 and we can compare them.

0:19:20.62 –> 0:19:23.86 And so instead we can sometimes try to get evidence on sort

0:19:23.86 –> 0:19:24.693 of the pieces.

0:19:24.693 –> 0:19:27.63 So, but again, we basically often have very little

0:19:27.63 –> 0:19:31.35 information on why people end up participating in trials.

0:19:31.35 –> 0:19:33.73 And we also are having,

0:19:33.73 –> 0:19:36.26 I think there's growing numbers of methods,

0:19:36.26 –> 0:19:38.57 but there's still limited information on treatment effect

0:19:38.57 –> 0:19:40.01 heterogeneity.

0:19:40.01 –> 0:19:42.57 Individual randomized trials are almost never powered

0:19:42.57 –> 0:19:45.24 to detect subgroup effects.

0:19:45.24 –> 0:19:47.76 Although, there is really growing research in this field

0:19:47.76 –> 0:19:50.193 and that is maybe a topic for another day.

0:19:52.38 –> 0:19:53.4 Okay.

0:19:53.4 –> 0:19:54.98 But again, there is a little. . .

0:19:54.98 –> 0:19:57.9 I think I'll go through this really quickly, but,

0:19:57.9 –> 0:20:01.11 I will give credit to some fields which are trying to better

0:20:01.11 –> 0:20:04.01 understand kind of who are the people that enroll in trials

0:20:04.01 –> 0:20:08.03 and how do they compare policy populations of interest.

0:20:08.03 –> 0:20:10.62 So a lot of that has been done in sort of the substance

0:20:10.62 –> 0:20:11.71 use field.

0:20:11.71 –> 0:20:14.24 And you can see a bunch of sites here

0:20:14.24 –> 0:20:17.97 documenting that people who participate in randomized trials

0:20:17.97 –> 0:20:21.76 of substance use treatment do actually differ quite

0:20:21.76 –> 0:20:25.05 substantially from people seeking treatment for substance

0:20:25.05 –> 0:20:26.88 use problems more generally.

0:20:26.88 –> 0:20:31.64 So for example, the Okuda reference the eligibility criteria

0:20:31.64 –> 0:20:35.51 in cannabis treatment RCTs would exclude about 80%

0:20:35.51 –> 0:20:38.16 of patients across the U.S. seeking treatment

0:20:38.16 –> 0:20:39.96 for cannabis use.

0:20:39.96 –> 0:20:42.9 And so again, it's sort of there's indications

0:20:42.9 –> 0:20:45.22 that the people that participate in trials

0:20:45.22 –> 0:20:47.9 are not necessarily reflective of the people

0:20:47.9 –> 0:20:50.183 for whom decisions are having to be made.

0:20:53.92 –> 0:20:57.42 Okay, so hopefully that at least kind of give some

0:20:57.42 –> 0:21:00.74 motivation for why we want to think more carefully

0:21:00.74 –> 0:21:03.63 about the population average treatment effect

0:21:03.63 –> 0:21:05.92 and why we might wanna think about designing studies

0:21:05.92 –> 0:21:09.67 or analyzing data in ways that help us estimate that.

0:21:09.67 –> 0:21:12.683 Any questions before I move to, how do we do that?

0:21:18.59 –> 0:21:19.91 Okay.

0:21:19.91 –> 0:21:21.09 I will end. . .

0:21:21.09 –> 0:21:24.37 I'm gonna hopefully end it at about 12:45, 1250,

0:21:24.37 –> 0:21:26.043 so we'll have time at the end, too.

0:21:27.461 –> 0:21:30.84 So, as a statistician, I feel obligated to say,

0:21:30.84 –> 0:21:32.27 and actually I have a quote on this at the very end

0:21:32.27 –> 0:21:33.42 of the talk.

0:21:33.42 –> 0:21:35.78 If we wanna be serious about estimating something,

0:21:35.78 –> 0:21:38.46 it's better to incorporate that through the design

0:21:38.46 –> 0:21:41.11 of our study, rather than trying to do it post talk

0:21:41.11 –> 0:21:41.943 at the end.

0:21:43.67 –> 0:21:46.73 So let's talk briefly about how we can improve external

0:21:46.73 –> 0:21:49.933 validity through study or randomized trial design.

0:21:51.687 –> 0:21:52.69 So again,

0:21:52.69 –> 0:21:55.99 as I alluded to earlier with the sort of ideal experiment.

0:21:55.99 –> 0:21:59.21 An ideal scenario is one where we can randomly sample

0:21:59.21 –> 0:22:02.48 from a population and then randomly assign treatment

0:22:02.48 –> 0:22:04.07 and control conditions.

0:22:04.07 –> 0:22:07.43 Doing this will give us a formerly unbiased treatment effect

0:22:07.43 –> 0:22:10.08 estimate in the population of interest.

0:22:10.08 –> 0:22:11.24 This is wonderful.

0:22:11.24 –> 0:22:14.703 I know of about six examples of this type.

0:22:16.96 –> 0:22:19.31 Most of the examples I know of are actually a federal

0:22:19.31 –> 0:22:22.66 government programs where they are administered through

0:22:22.66 –> 0:22:24.67 like centers or sites.

0:22:24.67 –> 0:22:27.96 And the federal government was able to mandate participation

0:22:27.96 –> 0:22:29.14 in an evaluation.

0:22:29.14 –> 0:22:32.75 So classic example is the Head Start Impact Study,

0:22:32.75 –> 0:22:36.42 where they were able to randomly select headstart centers

0:22:36.42 –> 0:22:37.26 to participate.

0:22:37.26 –> 0:22:39.26 And then within each center,

0:22:39.26 –> 0:22:42.29 they randomized kids to be able to get in off the wait list

0:22:42.29 –> 0:22:43.76 versus not.

0:22:43.76 –> 0:22:46.763 An upward bound evaluation had a very similar design.

0:22:47.73 –> 0:22:49.78 It's funny, I was. . .

0:22:49.78 –> 0:22:52.36 I gave a talk on this topic at Facebook and I was like,

0:22:52.36 –> 0:22:54.21 why is Facebook gonna care about this?

0:22:54.21 –> 0:22:56.1 Because you would think at a place like Facebook,

0:22:56.1 –> 0:22:58.54 they have their user sample,

0:22:58.54 –> 0:23:01.85 they should be able to do randomization within,

0:23:01.85 –> 0:23:04.18 like they should be able to pick users randomly

0:23:04.18 –> 0:23:06.36 and then do any sort of random assignment they want

0:23:06.36 –> 0:23:07.2 within that.

0:23:07.2 –> 0:23:10.27 It turns out it's more complicated than that, and so,

0:23:10.27 –> 0:23:12 they were interested in this topic,

0:23:12 –> 0:23:14.59 but I think that's another sort of example where people

0:23:14.59 –> 0:23:16.49 should be thinking, could we do this?

0:23:16.49 –> 0:23:17.52 Like,

0:23:17.52 –> 0:23:18.653 in a health system.

0:23:19.64 –> 0:23:22.39 I can imagine Geisinger or something implement something

0:23:22.39 –> 0:23:24.19 in their electronic health record where

0:23:24.19 –> 0:23:25.86 it's about messaging or something.

0:23:25.86 –> 0:23:29.02 And you could imagine actually picking people randomly

0:23:29.02 –> 0:23:30.6 to then randomize.

0:23:30.6 –> 0:23:32.1 But again, that's pretty rare.

0:23:33.14 –> 0:23:35.39 There's an idea that's called purpose of sampling.

0:23:35.39 –> 0:23:39.197 And this goes back to like the 1960s or 70s

0:23:39.197 –> 0:23:43.8 and the idea is sort of picking subjects purposefully.

0:23:43.8 –> 0:23:47.21 So one example here is like maybe we think

0:23:47.21 –> 0:23:49.33 that this intervention might look different

0:23:49.33 –> 0:23:51.76 or have different effects for large versus small

0:23:51.76 –> 0:23:52.593 school districts.

0:23:52.593 –> 0:23:55.75 So in our study, we just make an effort to enroll

0:23:55.75 –> 0:23:57.803 both large and small districts.

0:23:58.72 –> 0:23:59.63 This is sort of nice.

0:23:59.63 –> 0:24:04.373 It kind of gives you some variability in the types of people

0:24:05.41 –> 0:24:08.87 or subjects in the trial, but, it doesn't have the formal

0:24:08.87 –> 0:24:11.57 representativeness and sort of the formal unbiasness,

0:24:11.57 –> 0:24:14.51 like the random sampling I just talked about.

0:24:14.51 –> 0:24:17.21 And then again, sort of similar is this idea and this push

0:24:17.21 –> 0:24:20.06 in many fields towards pragmatic or practical clinical

0:24:20.06 –> 0:24:23.61 trials, where the idea is just to sort of try to enroll

0:24:23.61 –> 0:24:26.61 like kind of more representative sample

0:24:26.61 –> 0:24:28.78 in sort of a hand wavy way like I'm doing now.

0:24:28.78 –> 0:24:31.44 So not, it doesn't have this sort of formal statistical

0:24:31.44 –> 0:24:34.64 underpinning, but at least it's trying to make sure

0:24:34.64 –> 0:24:38.02 that it's not just patients from the Yale hospital

0:24:38.02 –> 0:24:41.12 and the Hopkins hospital and whatever sort of large medical

0:24:41.12 –> 0:24:44.51 centers, at least they might be trying to enroll patients

0:24:44.51 –> 0:24:46.703 from a broader spectrum across the U.S.

0:24:48.8 –> 0:24:52.97 Unfortunately, though, as much as I want to do things

0:24:52.97 –> 0:24:55.66 for design often, we're in a case where there's a study

0:24:55.66 –> 0:25:00.11 that's already been conducted and we are just

0:25:00.11 –> 0:25:01.31 sort of stuck analyzing it.

0:25:01.31 –> 0:25:04.42 And we wanna get a sense for how representative

0:25:04.42 –> 0:25:06.893 the results might be for a population.

0:25:08.74 –> 0:25:10.34 Sometimes people, when I talk about this,

0:25:10.34 –> 0:25:12.51 people are like, well, isn't this what meta-analysis does?

0:25:12.51 –> 0:25:16.08 Like meta-analysis enables you to combine multiple

0:25:16.08 –> 0:25:19.82 randomized trials and come up with sort of an overall

0:25:19.82 –> 0:25:20.723 effect estimate.

0:25:22.65 –> 0:25:26.41 And my answer to that is sort of yes maybe, or no maybe.

0:25:26.41 –> 0:25:29.65 Basically, the challenge with meta-analysis,

0:25:29.65 –> 0:25:33.76 is that until recently, no one really had a potential target

0:25:33.76 –> 0:25:35.27 population.

0:25:35.27 –> 0:25:38 It was not very formal about what the target population is.

0:25:38 –> 0:25:41.23 I think underlying that analysis is generally

0:25:41.23 –> 0:25:43.79 sort of a belief that the effects are constant

0:25:43.79 –> 0:25:45.793 and we're just trying to pool data.

0:25:47.538 –> 0:25:48.371 And it. . .

0:25:48.371 –> 0:25:49.76 And even just like, you can sort of see this,

0:25:49.76 –> 0:25:52.17 like if all of the trials sampled the same

0:25:52.17 –> 0:25:54.42 non-representative population,

0:25:54.42 –> 0:25:56.98 combining them is not going to help you get towards

0:25:56.98 –> 0:25:58.143 representativeness.

0:25:59.12 –> 0:26:01.41 That's that I have a former Postdoc Hwanhee Hong,

0:26:01.41 –> 0:26:02.85 who's now at Duke.

0:26:02.85 –> 0:26:05.54 And she has been doing some work to try to bridge

0:26:05.54 –> 0:26:07.97 these worlds and sort of really try to think through,

0:26:07.97 –> 0:26:11.59 well, how can we better use multiple trials

0:26:11.59 –> 0:26:14.233 to get to target population effects?

0:26:15.52 –> 0:26:18.34 There's another field it's called risk cross-design

0:26:18.34 –> 0:26:21.06 synthesis or research synthesis.

0:26:21.06 –> 0:26:22 This is sort of neat.

0:26:22 –> 0:26:26.17 It's one where you kind of combine randomized trial data,

0:26:26.17 –> 0:26:29.82 which might be not representative with non-experimental

0:26:29.82 –> 0:26:30.653 study data.

0:26:30.653 –> 0:26:34.32 So sort of explicitly trading off the internal and external

0:26:34.32 –> 0:26:35.93 validity.

0:26:35.93 –> 0:26:37.24 I'm not gonna get into the details,

0:26:37.24 –> 0:26:38.26 there's some references here.

0:26:38.26 –> 0:26:41.36 Ellie Kaizar at Ohio State, is one of the people

0:26:41.36 –> 0:26:43.283 that's done a lot of work on this.

0:26:45.31 –> 0:26:48.18 And part of the reason I'm not focused on this is that

0:26:48.18 –> 0:26:52.51 I work in a lot of areas like education and public health,

0:26:52.51 –> 0:26:54.05 sort of social science areas,

0:26:54.05 –> 0:26:56.18 where we often don't have multiple studies.

0:26:56.18 –> 0:27:00.47 So we often are stuck with just one study and we're trying

0:27:00.47 –> 0:27:03.97 to use that to learn about target populations.

0:27:03.97 –> 0:27:07.11 So I'm gonna briefly talk about an example

0:27:07.11 –> 0:27:11.81 where we trying to sort of do this.

0:27:11.81 –> 0:27:16.2 And basically, the fundamental idea is to re-weight

0:27:16.2 –> 0:27:19.563 the study sample to look like the target population.

0:27:20.78 –> 0:27:24.96 This idea is related to post stratification

0:27:24.96 –> 0:27:27.31 or, oh my gosh, I'm blanking now.

0:27:27.31 –> 0:27:29.423 Raking adjustments in surveys.

0:27:30.66 –> 0:27:33.49 So post stratification would be sort of at a simple level,

0:27:33.49 –> 0:27:34.74 would be something like. . .

0:27:34.74 –> 0:27:38.3 Well, if we know that males and females

0:27:38.3 –> 0:27:41.23 have different effects, or let's say young and old

0:27:41.23 –> 0:27:43.69 have different effects, let's estimate the effects

0:27:43.69 –> 0:27:46.153 separately for young versus old.

0:27:47.13 –> 0:27:50.86 And then re-weight those using the population proportions

0:27:50.86 –> 0:27:52.683 of sort of young versus old.

0:27:54.34 –> 0:27:57.55 That sort of stratification doesn't work if you have more

0:27:57.55 –> 0:28:02.45 than like one or two categorical effect moderators.

0:28:02.45 –> 0:28:03.283 And so,

0:28:03.283 –> 0:28:05.63 what I'm gonna show today is an approach where we use

0:28:05.63 –> 0:28:07.72 weighting, where we fit a model,

0:28:07.72 –> 0:28:10.08 predicting participation in the trial,

0:28:10.08 –> 0:28:13.1 and then weight the trial sample to look like the target

0:28:13.1 –> 0:28:14.1 population.

0:28:14.1 –> 0:28:16.96 So similar idea to things like propensity score weights

0:28:16.96 –> 0:28:20.253 or non-response adjustment weights in samples.

0:28:21.37 –> 0:28:23.15 There is a different approach,

0:28:23.15 –> 0:28:26.64 So what I'm gonna illustrate today is sort of this sample

0:28:26.64 –> 0:28:29.29 selection weighting strategy.

0:28:29.29 –> 0:28:32.07 You also can tackle this external validity

0:28:32.07 –> 0:28:34.88 by trying to model the outcome very flexibly

0:28:34.88 –> 0:28:39.013 and then project outcomes in the population.

0:28:40.45 –> 0:28:42.53 In some work I did with Jennifer Hill and others,

0:28:42.53 –> 0:28:45.52 we showed that BARTs, Bayesian Additive Regression Trees

0:28:45.52 –> 0:28:47.82 can actually work quite well for that purpose.

0:28:48.92 –> 0:28:52.58 And more recently, Issa Dahabreh at Brown has done some

0:28:52.58 –> 0:28:55.24 nice work sort of bridging these two and showing

0:28:55.24 –> 0:28:58.14 basically a doubly robust kind of idea where we can use

0:28:58.14 –> 0:29:03.14 both the sample membership model and the outcome model

0:29:03.58 –> 0:29:05.66 to have better performance.

0:29:05.66 –> 0:29:08.44 But today, I'm gonna just illustrate the weighting approach,

0:29:08.44 –> 0:29:10.7 partly because it's a really nice sort of pedagogical

0:29:10.7 –> 0:29:13.54 example and helps you kind of see what's going on

0:29:13.54 –> 0:29:14.373 in the data.

0:29:15.85 –> 0:29:18.373 Okay, any questions before I continue?

0:29:20.52 –> 0:29:21.353 Okay.

0:29:22.38 –> 0:29:25.67 So the example I'm gonna use is. . .

0:29:25.67 –> 0:29:28.08 There was this, I mean, some of you probably know much more

0:29:28.08 –> 0:29:32.53 about HIV treatment than I do, but the ACTG Trial,

0:29:32.53 –> 0:29:35.82 which was now quite an old trial,

0:29:35.82 –> 0:29:38.59 but it was one of the ones that basically showed that

0:29:38.59 –> 0:29:41.94 HAART therapy, highly active antiretroviral therapy

0:29:41.94 –> 0:29:46.19 was quite effective at reducing time to AIDS or death

0:29:46.19 –> 0:29:49.49 compared to standard combination therapy at the time.

0:29:49.49 –> 0:29:53.91 So it randomized about 1200 U.S. HIV positive adults

0:29:53.91 –> 0:29:56.44 to treatment versus control.

0:29:56.44 –> 0:29:59.38 And the intent to tree analysis in the trial

0:29:59.38 –> 0:30:01.46 had a hazard ratio of 0.51.

0:30:01.46 –> 0:30:05.513 So again, very effective at reducing time to AIDS or death.

0:30:06.87 –> 0:30:10.4 So Steve Cole and I though kind of asked the question, well,

0:30:10.4 –> 0:30:13.01 we don't necessarily just care about the people

0:30:13.01 –> 0:30:13.92 in the trial.

0:30:13.92 –> 0:30:16.49 This seems to be a very effective treatment.

0:30:16.49 –> 0:30:19.42 What could we use this data to project out

0:30:19.42 –> 0:30:21.83 sort of what the effects of the treatment would be

0:30:21.83 –> 0:30:24.53 if it were implemented nationwide?

0:30:24.53 –> 0:30:28.4 So we from CDC got estimates of the number of people

0:30:28.4 –> 0:30:31.92 newly infected with HIV in 2006.

0:30:31.92 –> 0:30:35.23 And basically, asked the question sort of if hypothetically,

0:30:35.23 –> 0:30:39.84 everyone in that group were able to get HAART versus

0:30:39.84 –> 0:30:41.67 standard combination therapy,

0:30:41.67 –> 0:30:44.833 what would be the population impacts of this treatment?

0:30:47.7 –> 0:30:50.33 In this case, because of sort of data availability,

0:30:50.33 –> 0:30:54.63 we only had the joint distribution of age, sex and race

0:30:54.63 –> 0:30:56.07 for the population.

0:30:56.07 –> 0:30:59.37 So we made sort of a pseudo population, again,

0:30:59.37 –> 0:31:01.5 sort of representing the U.S. population

0:31:01.5 –> 0:31:03.25 of newly infected people.

0:31:03.25 –> 0:31:05.78 But again, all we have is sex, race and age,

0:31:05.78 –> 0:31:07.08 which I will come back to.

0:31:08.49 –> 0:31:11.63 So this table documents the trial and the population.

0:31:11.63 –> 0:31:14.54 So you can see for example,

0:31:14.54 –> 0:31:19.54 that the trial tended to have more sort of 30 to 39 year

0:31:19.7 –> 0:31:23.773 olds, many fewer people under 30.

0:31:24.822 –> 0:31:28.6 The trial had more males and also had more whites

0:31:28.6 –> 0:31:32.28 and fewer blacks, Hispanic was similar.

0:31:32.28 –> 0:31:35.47 But I wanna flag and we'll come back to this in a minute

0:31:35.47 –> 0:31:37.85 that, in what I'm gonna show,

0:31:37.85 –> 0:31:41.15 we can adjust for the age, sex, race distribution.

0:31:41.15 –> 0:31:43 But, there's a real limitation,

0:31:43 –> 0:31:45.96 which is that the CD4 cell count as sort of a measure

0:31:45.96 –> 0:31:50.22 of disease severity is not available in the population.

0:31:50.22 –> 0:31:53.31 So this is a potential effect moderator,

0:31:53.31 –> 0:31:56.13 which we don't observe in the population.

0:31:56.13 –> 0:31:59.34 So in sort of projecting the impacts, we can say, well,

0:31:59.34 –> 0:32:02.74 here is the predicted impact given the age, sex,

0:32:02.74 –> 0:32:05.64 race distribution, but there's this unobserved

0:32:05.64 –> 0:32:09.37 potential effect moderator that we sort of might be worried

0:32:09.37 –> 0:32:11.32 about kind of in the back of our heads.

0:32:14.56 –> 0:32:16.52 So again, I briefly mentioned this,

0:32:16.52 –> 0:32:19.75 this is like the super basic description

0:32:19.75 –> 0:32:21.78 of what can be done.

0:32:21.78 –> 0:32:24.06 There are more nuances and I have some sites at the end

0:32:24.06 –> 0:32:25.89 for sort of more details.

0:32:25.89 –> 0:32:27.78 But basically fundamentally will, again,

0:32:27.78 –> 0:32:29.7 we sort of think about it as we kind of stack

0:32:29.7 –> 0:32:30.7 our data sets together.

0:32:30.7 –> 0:32:33.75 So we put our trial sample and our population data set

0:32:33.75 –> 0:32:34.75 together.

0:32:34.75 –> 0:32:37.94 We have an indicator for whether someone is in the trial

0:32:37.94 –> 0:32:39.69 versus the population.

0:32:39.69 –> 0:32:42.53 And then, we're gonna wait the trial members

0:32:42.53 –> 0:32:45.67 by their inverse probability of being in the trial

0:32:45.67 –> 0:32:48.47 as a function of the observed covariance.

0:32:48.47 –> 0:32:51.32 And again, very similar intuition and ideas

0:32:51.32 –> 0:32:54.65 and theory underlying this as underlying things

0:32:54.65 –> 0:32:57.63 like Horvitz-Thomson estimation in sample surveys

0:32:58.48 –> 0:33:00.68 and inverse probability of treatment waiting

0:33:00.68 –> 0:33:02.363 in non-experimental studies.

0:33:06.16 –> 0:33:09.31 So I showed you earlier that age, sex and race

0:33:09.31 –> 0:33:13.32 are all related to participation in the trial.

0:33:13.32 –> 0:33:15.45 What I'm not showing you the details of,

0:33:15.45 –> 0:33:18.5 but just trust me is that those factors also moderate

0:33:18.5 –> 0:33:20.465 effects in the trial.

0:33:20.465 –> 0:33:23.96 So the trial showed the largest effects for those ages,

0:33:23.96 –> 0:33:27.62 30 to 39, males and black individuals.

0:33:27.62 –> 0:33:30.62 And so, this is exactly why then what we might think

0:33:30.62 –> 0:33:34.15 that the overall trial estimate might not reflect

0:33:34.15 –> 0:33:36.383 what we would see population-wide.

0:33:38.72 –> 0:33:40.04 Ironically though, it turns out actually

0:33:40.04 –> 0:33:41.1 it kind of all cancels out.

0:33:41.1 –> 0:33:44.91 So this table shows the estimated population effects.

0:33:44.91 –> 0:33:48.05 So the first row again, is just the sort of naive trial

0:33:48.05 –> 0:33:49.66 results.

0:33:49.66 –> 0:33:52.39 We can then sort of weight by each characteristic

0:33:52.39 –> 0:33:55.7 separately, and then the bottom row is the combined

0:33:55.7 –> 0:33:57.86 age, sex, race adjustments.

0:33:57.86 –> 0:34:00.75 And you can see sort of actually the hazard ratio

0:34:00.75 –> 0:34:02.81 was remarkably similar.

0:34:02.81 –> 0:34:04.93 It's partly because like the age weightings

0:34:04.93 –> 0:34:07.1 sort of makes the impact smaller,

0:34:07.1 –> 0:34:09.61 but then the race weighting makes it bigger.

0:34:09.61 –> 0:34:11.56 And so then it kind of just washes out.

0:34:13.27 –> 0:34:14.59 But again, it's sort of a nice example,

0:34:14.59 –> 0:34:17.01 cause you can sort of see how the patterns

0:34:17.01 –> 0:34:19.9 evolve based on the size of the effects

0:34:19.9 –> 0:34:21.423 and the sample selection.

0:34:22.55 –> 0:34:24.77 I also wanna point out though that, of course,

0:34:24.77 –> 0:34:27.47 the confidence interval is wider,

0:34:27.47 –> 0:34:30.02 and that is sort of reflecting the fact that we are doing

0:34:30.02 –> 0:34:33.26 this extrapolation from the trial sample to the population.

0:34:33.26 –> 0:34:36.21 And so there's sort of a variance price we'll pay for that.

0:34:38.99 –> 0:34:39.823 Okay.

0:34:39.823 –> 0:34:43.61 So I haven't been super formal on the assumptions,

0:34:43.61 –> 0:34:45.11 but I'm I alluded to this?

0:34:45.11 –> 0:34:47.52 So I wanna just take a few minutes to turn

0:34:47.52 –> 0:34:50.1 to what about unobserved moderators?

0:34:50.1 –> 0:34:53.77 Because again, we can interpret this 0.57

0:34:53.77 –> 0:34:58.41 as the sort of overall population effect estimate

0:34:58.41 –> 0:35:01.42 only under an assumption that there are no unobserved

0:35:01.42 –> 0:35:05.55 moderators that differ between sample and population,

0:35:05.55 –> 0:35:08.063 once we adjust for age, sex, race.

0:35:11 –> 0:35:12.453 Okay, and in reality,

0:35:13.5 –> 0:35:16.61 such unobserved effect moderators are likely the rule,

0:35:16.61 –> 0:35:18.34 not the exception.

0:35:18.34 –> 0:35:20.41 So again, sort of, as I just said,

0:35:20.41 –> 0:35:23.11 the key assumption is that we've basically adjusted

0:35:23.11 –> 0:35:26.46 for all of the effect moderators.

0:35:26.46 –> 0:35:29.95 Very kind of comparable assumption to the assumption

0:35:29.95 –> 0:35:33.463 of no an observed confounding in a non-experimental study.

0:35:35.04 –> 0:35:37.9 And one of the reasons this is an important assumption

0:35:37.9 –> 0:35:41.69 to think about, is that, it is quite rare actually

0:35:41.69 –> 0:35:45.57 to have extensive covariate data overlap

0:35:45.57 –> 0:35:48.07 between the sample and the population.

0:35:48.07 –> 0:35:50.65 I have been working in this area for. . .

0:35:50.65 –> 0:35:51.69 How many years now?

0:35:51.69 –> 0:35:52.99 At least 10 years.

0:35:52.99 –> 0:35:55.83 And I've found time and time again,

0:35:55.83 –> 0:35:58.44 across a number of content areas,

0:35:58.44 –> 0:36:01.27 that it is quite rare to have a randomized trial sample

0:36:01.27 –> 0:36:03.38 and the target population dataset

0:36:03.38 –> 0:36:06.01 with very many comparable measures.

0:36:06.01 –> 0:36:07.82 So in the Stuart and Rhodes paper,

0:36:07.82 –> 0:36:11.52 this was in like early childhood setting

0:36:11.52 –> 0:36:15.33 and each data set, the trial and the population data

0:36:15.33 –> 0:36:19.35 had like over 400 variables observed at baseline.

0:36:19.35 –> 0:36:21.99 There were literally only seven that were measured

0:36:21.99 –> 0:36:24.63 consistently between the two samples.

0:36:24.63 –> 0:36:28.12 So essentially we have very limited ability then to adjust

0:36:28.12 –> 0:36:31.403 for these factors because they just don't have much overlap.

0:36:32.29 –> 0:36:37.02 So what that then motivated us to create some sensitivity

0:36:37.02 –> 0:36:40.11 analysis to basically probe and say, well,

0:36:40.11 –> 0:36:43.23 what if there is an unobserved effect moderator,

0:36:43.23 –> 0:36:47.16 how much would that change our population effect estimate?

0:36:47.16 –> 0:36:51.37 Again, this is very comparable to analysis of sensitivity,

0:36:51.37 –> 0:36:54.35 to unobserved confounding and non-experimental studies

0:36:54.35 –> 0:36:58.68 sort of adapted for this purpose of trial population,

0:36:58.68 –> 0:36:59.683 generalized ability.

0:37:03.22 –> 0:37:05.86 I think I can skip this in the interest of time and not go

0:37:05.86 –> 0:37:06.76 through all the details.

0:37:06.76 –> 0:37:08.22 If anyone wants the slides by the way,

0:37:08.22 –> 0:37:10.52 feel free to email me, I'm happy to send them.

0:37:12.8 –> 0:37:14.72 I'm gonna skip this too cause I've already said

0:37:14.72 –> 0:37:18.78 sort of the key assumption that is relevant for right now,

0:37:18.78 –> 0:37:22.333 but basically what we propose is,

0:37:23.802 –> 0:37:25.73 I'm gonna talk about two cases.

0:37:25.73 –> 0:37:29.37 So the easier case is this one where we're gonna assume

0:37:29.37 –> 0:37:32.28 that the randomized trial observes all of the effect

0:37:32.28 –> 0:37:33.113 moderators.

0:37:33.113 –> 0:37:36.35 And the issue is that our target population dataset

0:37:36.35 –> 0:37:40.62 does not have some moderators observed.

0:37:40.62 –> 0:37:43.1 I think this is fairly realistic because at least

0:37:43.1 –> 0:37:46.59 like to think that the people running the randomized trials

0:37:46.59 –> 0:37:49.52 have enough scientific knowledge and expertise

0:37:49.52 –> 0:37:52.39 that they sort of know what the likely effect moderators

0:37:52.39 –> 0:37:54.83 are and that they measure them in the trial.

0:37:54.83 –> 0:37:57.76 That is probably not fully realistic, but I'm. . .

0:37:57.76 –> 0:38:00.46 I like to give them sort of the benefit of the doubt

0:38:00.46 –> 0:38:01.47 on that.

0:38:01.47 –> 0:38:04.96 And that sort of that's what the ACTG example,

0:38:04.96 –> 0:38:07.47 was like CD4 count would be an example of this,

0:38:07.47 –> 0:38:10.84 where we have CD4 count in the trial,

0:38:10.84 –> 0:38:13.52 but we just don't have it in the population.

0:38:13.52 –> 0:38:16.06 So what we showed is that there's actually,

0:38:16.06 –> 0:38:18.06 a couple of different ways you can implement

0:38:18.06 –> 0:38:20.053 this sort of sensitivity analysis.

0:38:21.51 –> 0:38:24.6 One is essentially kind of an outcome model based one

0:38:24.6 –> 0:38:25.483 where you,

0:38:27.64 –> 0:38:30.32 basically, we just sort of specify a range

0:38:30.32 –> 0:38:34.15 for the unobserved moderator V in the population.

0:38:34.15 –> 0:38:36.27 So we kind of say, well, we don't know

0:38:36.27 –> 0:38:39.78 the distribution of this moderator in the population,

0:38:39.78 –> 0:38:43.01 but we're gonna guess that it's in some range.

0:38:43.01 –> 0:38:47.86 And then, we kind of projected out using data from the trial

0:38:47.86 –> 0:38:50.54 to understand like the extent of the moderation

0:38:50.54 –> 0:38:51.743 due to that variable.

0:38:52.9 –> 0:38:55.11 There's another variation on this,

0:38:55.11 –> 0:38:57.76 which is sort of the weighting variation

0:38:57.76 –> 0:38:59.92 where you kind of adjust the weights,

0:38:59.92 –> 0:39:03.43 essentially again for this unobserved moderator.

0:39:03.43 –> 0:39:07.15 Again, either way you sort of basically just have to specify

0:39:07.15 –> 0:39:11.44 a potential range for this V, the unobserved moderator

0:39:11.44 –> 0:39:12.593 in the population.

0:39:13.96 –> 0:39:15.603 So here's an example of that.

0:39:15.603 –> 0:39:18.28 This is a different example, where we were looking

0:39:18.28 –> 0:39:21.41 at the effects of a smoking cessation intervention

0:39:21.41 –> 0:39:24.46 among people in substance use treatment.

0:39:24.46 –> 0:39:29.46 And in the randomized trial, the mean addiction score

0:39:31.3 –> 0:39:33.03 was four.

0:39:33.03 –> 0:39:34.93 But we didn't have this addiction score,

0:39:34.93 –> 0:39:37.41 in the target population of interest.

0:39:37.41 –> 0:39:40.31 And so, what the sensitivity analysis allows us to do

0:39:40.31 –> 0:39:43.76 is to say, well, let's imagine that range is anywhere

0:39:43.76 –> 0:39:45.49 from three to five.

0:39:45.49 –> 0:39:49.1 And how much does that change our population effect

0:39:49.1 –> 0:39:50.52 estimates?

0:39:50.52 –> 0:39:53.52 Essentially, how steep this line is, is gonna be

0:39:53.52 –> 0:39:56.57 sort of determine how much it matters.

0:39:56.57 –> 0:39:58.8 And the steepness of the line basically

0:39:58.8 –> 0:40:01.72 is how much of a moderator is it,

0:40:01.72 –> 0:40:05.27 sort of how much effect heterogeneity is there in the trial

0:40:05.27 –> 0:40:07.49 as a result of that variable.

0:40:07.49 –> 0:40:10.58 But again, this is at least one way to sort of turn

0:40:10.58 –> 0:40:12.97 this sort of worry about an unobserved moderator

0:40:12.97 –> 0:40:15.77 into a more formal statement about how much

0:40:15.77 –> 0:40:17.083 it really might matter.

0:40:20.946 –> 0:40:22.39 I'm not gonna get into this partly,

0:40:22.39 –> 0:40:24.3 so you might also be thinking, well,

0:40:24.3 –> 0:40:27.367 what if the trial doesn't know what all the moderators are?

0:40:27.367 –> 0:40:30.6 And what if there's some fully unobserved moderator

0:40:30.6 –> 0:40:31.773 that will call U?

0:40:33.62 –> 0:40:35.65 This is a much much harder, basically,

0:40:35.65 –> 0:40:38.688 if anyone wants to try to dig into it, that would be great.

0:40:38.688 –> 0:40:41.66 Part of the reason it's harder is because you have to make

0:40:41.66 –> 0:40:44.38 very strong assumptions about the distribution

0:40:44.38 –> 0:40:47.99 of the observed covariance and U together.

0:40:47.99 –> 0:40:49.12 We put out one approach,

0:40:49.12 –> 0:40:52.92 but it is a fairly special case and not very general.

0:40:52.92 –> 0:40:56.03 So again, hopefully we're not in this sort of scenario

0:40:56.03 –> 0:40:56.863 very often.

0:41:00.59 –> 0:41:02.56 This is a little bit of a technicality,

0:41:02.56 –> 0:41:05.33 but often epidemiologists ask this question.

0:41:05.33 –> 0:41:08.63 So I've laid stuff out again with respect to kind of a risk

0:41:08.63 –> 0:41:10.53 difference or a difference in outcomes

0:41:11.64 –> 0:41:15.09 and sort of like more of like an additive treatment scale.

0:41:15.09 –> 0:41:17.41 There is this real complication that arises,

0:41:17.41 –> 0:41:19.98 which is that if you have like a binary,

0:41:19.98 –> 0:41:24.153 like the scale of the outcome matters in terms of effect

0:41:25.16 –> 0:41:26.32 moderation.

0:41:26.32 –> 0:41:29.56 And in particular, there might be sort of more apparent

0:41:29.56 –> 0:41:32.97 effect heterogeneity on one scale versus another.

0:41:32.97 –> 0:41:36.72 So I'm just kind of flagging this, that like this exists,

0:41:36.72 –> 0:41:39 there are some people sort of looking at this in more

0:41:39 –> 0:41:44 formal, but again for now sort of just think about like risk

0:41:44.16 –> 0:41:45.41 difference kind of scale.

0:41:47.45 –> 0:41:48.283 Okay, great.

0:41:48.283 –> 0:41:51.4 So let me just conclude with a few kind of final thoughts.

0:41:51.4 –> 0:41:54.44 So, I think all of us, not all of us,

0:41:54.44 –> 0:41:57.61 but often we sort of want to assume that study results

0:41:57.61 –> 0:41:58.443 generalize.

0:41:58.443 –> 0:42:01.13 Often people write a discussion section in a paper,

0:42:01.13 –> 0:42:04.56 where they kind of qualitatively have some sentences

0:42:04.56 –> 0:42:07.83 about why they do or don't think that the results

0:42:07.83 –> 0:42:10.19 in this paper kind of extend to other groups

0:42:10.19 –> 0:42:11.403 or other populations.

0:42:12.52 –> 0:42:16.18 But I think until the past again, sort of five or so years,

0:42:16.18 –> 0:42:19.14 a lot of that discussion was very hand-wavy

0:42:19.14 –> 0:42:20.81 and sort of qualitative.

0:42:20.81 –> 0:42:23.54 I think that what we are seeing in epidemiology

0:42:23.54 –> 0:42:26.07 and statistics and bias statistics

0:42:26.07 –> 0:42:29 recently has been a push towards having more

0:42:29 –> 0:42:33.16 ability to quantify this and make it sort of more formal

0:42:33.16 –> 0:42:33.993 statements.

0:42:35.04 –> 0:42:37.44 So I think if we do wanna be serious though,

0:42:37.44 –> 0:42:40.59 about assessing and enhancing external validity,

0:42:40.59 –> 0:42:42.6 again, we really need these different pieces.

0:42:42.6 –> 0:42:46.04 We need information on the factors that influence effect

0:42:46.04 –> 0:42:48.54 heterogeneity the moderators.

0:42:48.54 –> 0:42:50.7 We need information on the factors that influence

0:42:50.7 –> 0:42:54.86 participation in rigorous studies like randomized trials.

0:42:54.86 –> 0:42:57.37 And we need data on all of those things,

0:42:57.37 –> 0:42:59.173 in the trial and the population.

0:43:00.38 –> 0:43:03.5 And then finally, we need statistical methods that allow us

0:43:03.5 –> 0:43:07.103 to use that data to estimate population treatment effects.

0:43:07.94 –> 0:43:11.9 I would argue that that last bullet is sort of much further

0:43:11.9 –> 0:43:13.43 along than any of the others.

0:43:13.43 –> 0:43:15.49 That in my experience,

0:43:15.49 –> 0:43:18.7 the limiting factor is usually not the methods.

0:43:18.7 –> 0:43:22.23 The limiting factor at this point in time is the data

0:43:22.23 –> 0:43:24.61 and sort of the scientific knowledge

0:43:24.61 –> 0:43:27.033 about these different factors.

0:43:29.05 –> 0:43:30.24 And that's what this slide is.

0:43:30.24 –> 0:43:32.64 So I think I've already said, but that again,

0:43:32.64 –> 0:43:35.45 is sort of one of the motivations for the sensitivity

0:43:35.45 –> 0:43:38.87 analysis is just a recognition that it's often,

0:43:38.87 –> 0:43:40.84 really quite hard to get data that

0:43:42.02 –> 0:43:45.193 is consistently measured between a trial and a population.

0:43:46.71 –> 0:43:48.73 So on that point, recommendations again,

0:43:48.73 –> 0:43:51.34 if we wanna be serious about effect heterogeneity

0:43:51.34 –> 0:43:54.78 or about estimating population treatment effects,

0:43:54.78 –> 0:43:58.17 we need better information on treatment effect heterogeneity

0:43:59.21 –> 0:44:01.69 that might be better analysis of existing trials,

0:44:01.69 –> 0:44:04.5 that might be meta-analysis of existing trials.

0:44:04.5 –> 0:44:07.44 That might also be theoretical models for the interventions

0:44:07.44 –> 0:44:10.773 to understand what the likely moderators are.

0:44:11.83 –> 0:44:14.04 We also need better information on the factors

0:44:14.04 –> 0:44:17.16 that influence participation in trials and more discussion

0:44:17.16 –> 0:44:19.913 of how trial samples are selected.

0:44:21.86 –> 0:44:23.33 We need to standardize measures.

0:44:23.33 –> 0:44:26.25 So again, it's incredibly frustrating when you have trial

0:44:26.25 –> 0:44:29.66 and population data, but the measures in them are not

0:44:29.66 –> 0:44:30.89 consistent.

0:44:30.89 –> 0:44:33.44 There are methods that can be used for this,

0:44:33.44 –> 0:44:35.453 some data harmonization approaches,

0:44:36.39 –> 0:44:38.86 but, they require assumptions.

0:44:38.86 –> 0:44:42.45 It's better if we can be thoughtful and strategic about,

0:44:42.45 –> 0:44:45.25 for example, common measures across studies.

0:44:45.25 –> 0:44:47.07 I will say one of the frustrations too,

0:44:47.07 –> 0:44:50.83 is that in some fields like the early childhood data

0:44:50.83 –> 0:44:52.07 I talked about,

0:44:52.07 –> 0:44:54.56 part of the problem was like the two data sets might

0:44:54.56 –> 0:44:56.44 actually have the same measure,

0:44:56.44 –> 0:44:58.41 but they didn't give the raw data,

0:44:58.41 –> 0:45:00.63 and they're like standardized scales differently.

0:45:00.63 –> 0:45:03.3 Like they standardized them to their own population,

0:45:03.3 –> 0:45:04.79 not sort of more generally.

0:45:04.79 –> 0:45:08.343 And so they, weren't sort of on the same scale in the end.

0:45:09.9 –> 0:45:12.26 As a statistician, of course, I will say we do need more

0:45:12.26 –> 0:45:15.26 research on the methods and understanding when they work

0:45:15.26 –> 0:45:16.093 and when they don't.

0:45:16.093 –> 0:45:18.63 There are some pretty strong assumptions

0:45:18.63 –> 0:45:20.35 in these approaches.

0:45:20.35 –> 0:45:23.84 But again, I think that sort of in some ways,

0:45:23.84 –> 0:45:26.893 that is further along and then some of the data situations.

0:45:28.68 –> 0:45:31.76 So I just wanted to take one minute to flag some current

0:45:31.76 –> 0:45:34.46 work in case partly if anyone wants to ask questions about

0:45:34.46 –> 0:45:36.11 these.

0:45:36.11 –> 0:45:38.22 One thing I'm kind of excited about,

0:45:38.22 –> 0:45:41.5 especially in my education world is. . .

0:45:41.5 –> 0:45:43.67 So what I've been talking about today has mostly been,

0:45:43.67 –> 0:45:46.01 if we have a trial sample and we wanna project

0:45:46.01 –> 0:45:48.73 to kind of a larger target population.

0:45:48.73 –> 0:45:50.71 But there's an equally interesting question,

0:45:50.71 –> 0:45:54.18 which is sort of how well can randomized trial informs

0:45:54.18 –> 0:45:55.61 or local decision making?

0:45:55.61 –> 0:46:00.043 So if we have a randomized trial with 60 schools in it,

0:46:00.99 –> 0:46:04.48 how well can the results from that trial be used to inform

0:46:04.48 –> 0:46:06.91 individual school districts decisions?

0:46:06.91 –> 0:46:08.892 Turns out, not particularly well.

0:46:08.892 –> 0:46:10 (laughs)

0:46:10 –> 0:46:11.92 We can talk more about that.

0:46:11.92 –> 0:46:15.04 I mentioned earlier, Issa Dahabreh, who's at Brown,

0:46:15.04 –> 0:46:18.1 and he's really interested in developing sort of the formal

0:46:18.1 –> 0:46:20.94 theories underlying different ways of estimating

0:46:20.94 –> 0:46:23.44 these population effects, again, including some

0:46:23.44 –> 0:46:25.163 doubly robust approaches.

0:46:26.368 –> 0:46:29.13 Trang Nguyen, who works at Hopkins with me,

0:46:29.13 –> 0:46:31.65 we are still looking at sort of the sensitivity analysis

0:46:31.65 –> 0:46:34.09 for unobserved moderators.

0:46:34.09 –> 0:46:37.19 I mentioned Hwanhee Hong already, who's now at Duke.

0:46:37.19 –> 0:46:40.45 And she, again, sort of straddles the meta-analysis world

0:46:40.45 –> 0:46:43 in this world, which has some really interesting

0:46:43 –> 0:46:43.833 connections.

0:46:44.91 –> 0:46:47.64 My former student now he's at Flatiron Health

0:46:47.64 –> 0:46:49.56 as of a few months ago.

0:46:49.56 –> 0:46:53.04 Ben Ackerman, did some work on sort of measurement error

0:46:53.04 –> 0:46:55.25 and sort of partly how to deal with some of these

0:46:55.25 –> 0:46:58.793 measurement challenges between the sample and population.

0:46:59.776 –> 0:47:03.58 And then I'll just briefly mention Daniel Westreich at UNC,

0:47:03.58 –> 0:47:05.04 who is really. . .

0:47:05.04 –> 0:47:08.7 If you come from sort of more of an epidemiology world,

0:47:08.7 –> 0:47:11.12 Daniel has some really nice papers that are sort of trying

0:47:11.12 –> 0:47:14.3 to translate these ideas to epidemiology,

0:47:14.3 –> 0:47:17.32 and this concept of what he calls target validity.

0:47:17.32 –> 0:47:20.25 So sort of rather than thinking about internal and external

0:47:20.25 –> 0:47:23.22 validity separately, and as potentially,

0:47:23.22 –> 0:47:25.69 in kind of conflict with each other,

0:47:25.69 –> 0:47:28.63 instead really think carefully about a target of inference

0:47:28.63 –> 0:47:31.22 and then thinking of internal and external validity

0:47:31.22 –> 0:47:34.83 sort of within that and not sort of trying to prioritize

0:47:34.83 –> 0:47:35.993 one over the other.

0:47:37.18 –> 0:47:39.133 And then just an aside, one thing,

0:47:39.981 –> 0:47:42.61 I would love to do more in the coming years is thinking

0:47:42.61 –> 0:47:45.58 about combining experimental and non-experimental evidence.

0:47:45.58 –> 0:47:48.66 I think that is probably where it would be very beneficial

0:47:48.66 –> 0:47:51.78 to go instead of more of that cross designed synthesis

0:47:51.78 –> 0:47:53.083 kind of idea.

0:47:54.81 –> 0:47:57.35 But again, I wanna conclude with this,

0:47:57.35 –> 0:48:00.95 which is gets us back to design and that again,

0:48:00.95 –> 0:48:04.04 sort of what is often the limiting factor here is the data

0:48:04.04 –> 0:48:06.96 and just sort of strong designs.

0:48:06.96 –> 0:48:10.13 So Rubin, 2005 with better data, fewer assumptions

0:48:10.13 –> 0:48:12.98 are needed and then Light, Singer and Willett,

0:48:12.98 –> 0:48:15.68 who are sort of big education methodologists.

0:48:15.68 –> 0:48:19.46 You can't fix by analysis what you've bungled by design.

0:48:19.46 –> 0:48:21.97 So again, just wanna highlight that if we wanna be serious

0:48:21.97 –> 0:48:24.42 about estimating population effects,

0:48:24.42 –> 0:48:26.99 we need to be serious about that in our study designs,

0:48:26.99 –> 0:48:29.61 both in terms of who we recruit,

0:48:29.61 –> 0:48:32.157 but then also what variables we collect on them.

0:48:32.157 –> 0:48:33.07 But if we do that,

0:48:33.07 –> 0:48:36.73 I think that we can have the potential to really help guide

0:48:36.73 –> 0:48:39.38 policy and practice by thinking more carefully

0:48:39.38 –> 0:48:41.843 about the populations that we care about.

0:48:43.02 –> 0:48:44.33 So for more. . .

0:48:44.33 –> 0:48:46.6 Here's this, there's my email, if you wanna email me

0:48:46.6 –> 0:48:48.5 for the slides.

0:48:48.5 –> 0:48:52.67 And thanks to various funders, and then I'll leave this up

0:48:52.67 –> 0:48:54.56 for a couple minutes,

0:48:54.56 –> 0:48:58.75 which are all big, tiny font, some of the references,

0:48:58.75 –> 0:49:01.06 but then I'll take that down in a minute so that we can see

0:49:01.06 –> 0:49:01.893 each other more.

0:49:01.893 –> 0:49:05.973 So thank you, and I'm very happy to take some questions.

0:49:13.78 –> 0:49:15.5 I don't know if you all have a way to organize

0:49:15.5 –> 0:49:16.4 or people just can

0:49:18.99 –> 0:49:19.823 jump in.

0:49:24.16 –> 0:49:25.2 - So maybe I'll ask the question.

0:49:25.2 –> 0:49:28.003 Thanks Liz, for this very interesting and great talk.

0:49:29.03 –> 0:49:33.5 So I noticed that you've talked about the target population

0:49:33.5 –> 0:49:34.89 in this framework.

0:49:34.89 –> 0:49:39.27 And I think there are situations where the population sample

0:49:39.27 –> 0:49:42.774 is actually a survey from a larger population.

0:49:42.774 –> 0:49:43.607 - Yeah.

0:49:43.607 –> 0:49:46.63 - Cause we do not really afford to absorb everything,

0:49:46.63 –> 0:49:48.75 actual population, which will contain

0:49:48.75 –> 0:49:50.11 like millions of individuals.

0:49:50.11 –> 0:49:54.83 And so in that situation, does the framework still apply

0:49:54.83 –> 0:49:58.37 particularly in terms of the sensitivity analysis?

0:49:58.37 –> 0:50:01.36 And is there any caveat that we should also know in dealing

0:50:01.36 –> 0:50:02.293 with those data?

0:50:03.33 –> 0:50:04.223 - Great question.

0:50:05.15 –> 0:50:07.24 And actually, thank you for asking that because I forgot

0:50:07.24 –> 0:50:09.6 to mention that Ben Ackerman's dissertation,

0:50:09.6 –> 0:50:10.5 also looked at that.

0:50:10.5 –> 0:50:12.92 So I mentioned his measurement error stuff.

0:50:12.92 –> 0:50:16.9 But yes, actually, so Ben's second dissertation paper

0:50:16.9 –> 0:50:20.95 did exactly that, where we sort of laid out the theory

0:50:20.95 –> 0:50:24.1 for when these the target population data

0:50:24.1 –> 0:50:27.033 comes from a complex survey itself.

0:50:28.65 –> 0:50:30.88 Short answer is yes, it all still works.

0:50:30.88 –> 0:50:34.46 Like you have to use the weights, there are some nuances,

0:50:34.46 –> 0:50:36.45 but, and you're right, like essentially,

0:50:36.45 –> 0:50:38.45 especially like in...

0:50:38.45 –> 0:50:41.31 Like for representing the U.S. population, often, the data

0:50:41.31 –> 0:50:44.29 we have is like the National Health Interview Survey

0:50:44.29 –> 0:50:47.04 or the Add Health Survey of Adolescents,

0:50:47.04 –> 0:50:49.11 which are these complex surveys.

0:50:49.11 –> 0:50:52.76 So short answer is, yeah, it still can work.

0:50:52.76 –> 0:50:54.943 Your question about the sensitivity analysis is actually

0:50:54.943 –> 0:50:57.9 a really good one and we have not extended...

0:50:57.9 –> 0:50:59.72 I'd have to think, I don't know, off hand, like,

0:50:59.72 –> 0:51:03.84 I think it would be sort of straightforward to extend

0:51:03.84 –> 0:51:06.56 the sensitivity analysis to that, but we haven't actually

0:51:06.56 –> 0:51:07.393 done it.

0:51:08.34 –> 0:51:09.173 - Thanks Liz.

0:51:10.73 –> 0:51:12.27 The other short question is that I noticed that

0:51:12.27 –> 0:51:16.38 in your slide, you first define, PATE as population ate,

0:51:16.38 –> 0:51:18.65 but then in one slide you have this Tate,

0:51:18.65 –> 0:51:21.15 which I assume is target ate.

0:51:21.15 –> 0:51:24.57 And so, I'm just really curious as to like, is there any,

0:51:24.57 –> 0:51:26.878 like differences or nuances in the choice of this

0:51:26.878 –> 0:51:27.943 terminology?

0:51:28.977 –> 0:51:29.81 - Good question.

0:51:29.81 –> 0:51:30.643 And no, yeah, I'm not...

0:51:30.643 –> 0:51:33.563 I wasn't very precise with that, but in my mind, no.

0:51:34.75 –> 0:51:37.83 Over time I've been trying to use Tate,

0:51:37.83 –> 0:51:39.97 but you can see that kind of just by default,

0:51:39.97 –> 0:51:41.713 I still sometimes use PATE.

0:51:42.83 –> 0:51:45.75 Part of the reason I use Tate is because I think

0:51:45.75 –> 0:51:48.02 the target is just a slightly more general term.

0:51:48.02 –> 0:51:50.21 Like people sometimes I think, think if we meet,

0:51:50.21 –> 0:51:53.33 if we say PATE, the population has to be like

0:51:53.33 –> 0:51:58.03 the U.S. population or some like very sort of big,

0:51:58.03 –> 0:52:00.93 very official population in some sense.

0:52:00.93 –> 0:52:03.57 Whereas, the target average treatment effect,

0:52:03.57 –> 0:52:06.26 Tate terminology, I think reflects that sometimes

0:52:06.26 –> 0:52:10.06 it's just a target group that's well-defined.

0:52:10.06 –> 0:52:10.893 - Gotcha.

0:52:10.893 –> 0:52:12.27 Thanks, that's very helpful.

0:52:12.27 –> 0:52:14.93 And I think we have a question coming from the chat as well.

0:52:14.93 –> 0:52:15.9 - Yeah, I just saw that.

0:52:15.9 –> 0:52:17.45 So I can read that.

0:52:17.45 –> 0:52:19.61 We have theory for inference from a sample to a target

0:52:19.61 –> 0:52:22.7 population needs to find that internal validity approaches,

0:52:22.7 –> 0:52:25.21 what theory is there for connecting the internal validity

0:52:25.21 –> 0:52:26.933 methods to external validity?

0:52:28.62 –> 0:52:32.55 So I think, what you mean is sort of,

0:52:32.55 –> 0:52:36.5 what is the formal theory for projecting the impact

0:52:36.5 –> 0:52:38.11 to the target population?

0:52:38.11 –> 0:52:40.7 That is exactly what some of those people that I referenced

0:52:40.7 –> 0:52:41.533 sort of lay out.

0:52:41.533 –> 0:52:42.366 Like I didn't. . .

0:52:42.366 –> 0:52:44.59 For this talk, I didn't get into all the theoretical weeds,

0:52:44.59 –> 0:52:46.37 but if you're interested in that stuff,

0:52:46.37 –> 0:52:48.83 probably some of Issa Dahabreh's work would be the most

0:52:48.83 –> 0:52:50.093 relevant to look at.

0:52:51.43 –> 0:52:54 Cause he really lays out sort of the formal theory.

0:52:54 –> 0:52:58.39 I mean, some of my early papers on this topic did it,

0:52:58.39 –> 0:53:01.22 but his is like a little bit more formal and sort of makes

0:53:01.22 –> 0:53:03.61 connections to the doubly robust literature

0:53:03.61 –> 0:53:04.443 and things like that.

0:53:04.443 –> 0:53:06.04 And so it's really...

0:53:06.04 –> 0:53:08.42 Anyway, that's what this whole literature

0:53:08.42 –> 0:53:11.05 and part of it is sort of building is that theoretical base

0:53:11.05 –> 0:53:12.223 for doing this.

0:53:17.32 –> 0:53:18.503 Any other questions?

0:53:28.07 –> 0:53:28.903 - [Ofer] Liz,

0:53:28.903 –> 0:53:30.226 I'm Ofer Harel.

0:53:30.226 –> 0:53:31.36 - Oh, hi Ofer?

0:53:31.36 –> 0:53:32.67 - [Ofer] Hi.

0:53:32.67 –> 0:53:33.63 (mumbles)

0:53:33.63 –> 0:53:37.453 Just jump on the corridor, so it's make it great.

0:53:39.01 –> 0:53:43.07 So in most of the studies that I would work on,

0:53:43.07 –> 0:53:45.86 they don't do really have a great idea about

0:53:45.86 –> 0:53:50.1 what really the population is and how really to measure

0:53:50.1 –> 0:53:50.933 those.

0:53:50.933 –> 0:53:53.59 So it's great if I have some measure of the population,

0:53:53.59 –> 0:53:57.41 but most of the time it is the studies that I work.

0:53:57.41 –> 0:54:01.63 I have no real measurements on that population.

0:54:01.63 –> 0:54:03.06 What happens then?

0:54:03.06 –> 0:54:03.977 - Yeah, great question.

0:54:03.977 –> 0:54:05.65 And in part, I meant to say this,

0:54:05.65 –> 0:54:07.5 but that's one of the reasons why the analogy...

0:54:07.5 –> 0:54:10.3 Why the design strategies don't always work particularly

0:54:10.3 –> 0:54:12.69 well is like, especially when you're just starting out

0:54:12.69 –> 0:54:13.523 a study, right?

0:54:13.523 –> 0:54:15.973 We don't really know the target population.

0:54:17.07 –> 0:54:21.28 I think certainly to do any of these procedures,

0:54:21.28 –> 0:54:24.84 you need eventually to have a well defined population.

0:54:24.84 –> 0:54:26.95 But I think that's partly why some of the analysis

0:54:26.95 –> 0:54:28.9 approaches are useful is that,

0:54:28.9 –> 0:54:31.09 you might have multiple target populations.

0:54:31.09 –> 0:54:33.01 Like we might have one trial,

0:54:33.01 –> 0:54:35.21 and we might be interested in saying,

0:54:35.21 –> 0:54:38.67 how well does this generalize to the State of New Hampshire

0:54:38.67 –> 0:54:41.37 or the State of Vermont or the State of Connecticut?

0:54:41.37 –> 0:54:45.32 And so, you could imagine one study that's used to inform

0:54:45.32 –> 0:54:47.103 multiple target populations.

0:54:48.05 –> 0:54:49.03 With different assumptions,

0:54:49.03 –> 0:54:50.47 sort of you have to think through the assumptions

0:54:50.47 –> 0:54:51.323 for each one.

0:54:52.39 –> 0:54:53.62 If you don't even,

0:54:53.62 –> 0:54:55.65 I guess I would say if you don't even know

0:54:55.65 –> 0:54:58.56 who your population is, you shouldn't be using these methods

0:54:58.56 –> 0:55:02.04 at all, cause like the whole premise is that there is some

0:55:02.04 –> 0:55:04.9 well-defined target population and you do need data on it

0:55:04.9 –> 0:55:05.93 or at least. . .

0:55:06.99 –> 0:55:09.34 Yeah, the joint distribution of some covariance

0:55:09.34 –> 0:55:10.38 or something.

0:55:10.38 –> 0:55:13.48 Without that, you're kind of just like,

0:55:13.48 –> 0:55:14.97 I don't know, what a good analogy is,

0:55:14.97 –> 0:55:17.923 but you're kinda just like guessing at everything.

0:55:23.936 –> 0:55:25.65 (mumbles)

0:55:25.65 –> 0:55:27.246 - No, go ahead.

0:55:27.246 –> 0:55:28.864 Go ahead.

0:55:28.864 –> 0:55:30.297 - Oh, Vinod, yeah.

0:55:30.297 –> 0:55:32.38 All my friends are popping up, it's great.

0:55:32.38 –> 0:55:34.37 (laughs)

0:55:34.37 –> 0:55:35.203 - [Vinod] Can I go ahead?

0:55:35.203 –> 0:55:36.923 I feel like I'm talking to someone.

0:55:38.66 –> 0:55:39.98 - Yeah, go ahead Vinod.

0:55:39.98 –> 0:55:42.1 - [Vinod] That was a great talk.

0:55:42.1 –> 0:55:44.32 So I have a little ill formulated question,

0:55:44.32 –> 0:55:47.13 but it's queuing after just the last question

0:55:47.13 –> 0:55:48.956 that was asked is,

0:55:48.956 –> 0:55:53.773 in clinical set populations where,

0:55:54.85 –> 0:55:57.62 in some ways we're using this clinical samples

0:55:57.62 –> 0:56:01.55 to learn about the population because unless they seek help,

0:56:01.55 –> 0:56:05.32 we often don't know what they are in the wild, so to speak.

0:56:05.32 –> 0:56:09.41 And so, each sampling of that clinical population

0:56:09.41 –> 0:56:12.84 is a maybe by sampling of that larger population

0:56:12.84 –> 0:56:14.1 in the wild.

0:56:14.1 –> 0:56:18.45 So I guess my question is, how do you get around this,

0:56:18.45 –> 0:56:21.73 I guess Rumsfeld problem, which is every time you sample

0:56:21.73 –> 0:56:24.14 there's this unknown, unknown, but there's no way to get

0:56:24.14 –> 0:56:27.34 at them because in some ways, your sampling relies on...

0:56:27.34 –> 0:56:29.85 If we could say it relies on help seeking,

0:56:29.85 –> 0:56:33.21 which is by itself as process.

0:56:33.21 –> 0:56:35.16 And if we could just stipulate, there's no way to get

0:56:35.16 –> 0:56:36.27 around that.

0:56:36.27 –> 0:56:38.653 How do you see this going forward?

0:56:39.55 –> 0:56:40.383 - Yeah, good question.

0:56:40.383 –> 0:56:42.65 I think right, particularly relevant in mental health

0:56:42.65 –> 0:56:45.68 research where there's a lot of people who are not seeking

0:56:45.68 –> 0:56:47.106 treatment.

0:56:47.106 –> 0:56:50.09 These methods are not gonna help with that in a sense

0:56:50.09 –> 0:56:53.09 like again, they are gonna be sort of tuned to whatever

0:56:53.09 –> 0:56:54.96 population you have.

0:56:54.96 –> 0:56:56.8 I think though there are. . .

0:56:56.8 –> 0:56:59.513 If you really wanna be thoughtful about that's

0:57:00.42 –> 0:57:02.87 problem, that's where sort of some of the strategies

0:57:02.87 –> 0:57:05.38 that were used like the Epidemiologic Catchment Area

0:57:05.38 –> 0:57:08.32 Surveys, where they would go door to door and knock on doors

0:57:08.32 –> 0:57:10.66 and do diagnostic interviews.

0:57:10.66 –> 0:57:14.07 Like if we wanna be really serious about trying to reach

0:57:14.07 –> 0:57:16.73 everyone and get an estimate of the really sort of true

0:57:16.73 –> 0:57:20.08 population, then we really have to tackle that

0:57:20.08 –> 0:57:23.253 very creatively and with a lot of resources probably.

0:57:25.027 –> 0:57:26.995 - [Vinod] Thanks.

0:57:26.995 –> 0:57:27.828 - Welcome.

0:57:29.15 –> 0:57:30.43 - Hi Liz?

0:57:30.43 –> 0:57:32.96 Yeah, it's gonna be a true question and great talk

0:57:32.96 –> 0:57:33.793 by the way.

0:57:34.91 –> 0:57:37.576 I'm curious, you mentioned there could be a slight

0:57:37.576 –> 0:57:40.189 difference between the terms transportability

0:57:40.189 –> 0:57:41.07 and generalizability.

0:57:41.07 –> 0:57:42.91 Yeah, I'm curious about that.

0:57:42.91 –> 0:57:45.91 - Yeah, briefly, this is a little bit of a. . .

0:57:47.563 –> 0:57:48.396 What's the word?

0:57:48.396 –> 0:57:51.12 Simplification, but briefly I think of generalizability

0:57:51.12 –> 0:57:54.67 as one where the sample that, like the trial sample

0:57:54.67 –> 0:57:57.12 is a proper subset of the population.

0:57:57.12 –> 0:58:01.46 So we do a trial in New Hampshire,

0:58:01.46 –> 0:58:04.18 and we're trying to generalize to new England.

0:58:04.18 –> 0:58:07.58 Whereas transportability is one where it is not a proper

0:58:07.58 –> 0:58:10.27 subset, so we do a trial in the United States

0:58:10.27 –> 0:58:12.143 and we wanna transport to Europe.

0:58:13.53 –> 0:58:16.69 Underlying both, the reason I don't worry too much about it,

0:58:16.69 –> 0:58:18.725 the terms is because either way,

0:58:18.725 –> 0:58:20.76 the assumption is essentially the same.

0:58:20.76 –> 0:58:23.13 Like you still have to make this assumption about

0:58:23.13 –> 0:58:25.11 no unobserved moderators.

0:58:25.11 –> 0:58:27.68 It's just that it's probably gonna be a stronger assumption

0:58:27.68 –> 0:58:29.544 and harder to believe,

0:58:29.544 –> 0:58:33.4 when transporting rather than when generalizing.

0:58:33.4 –> 0:58:36.47 Cause you sort of know that you're going from one place

0:58:36.47 –> 0:58:38.053 to another in some sense.

0:58:39.38 –> 0:58:40.5 - Thanks, makes sense.

0:58:40.5 –> 0:58:41.333 - Sure.

0:58:42.56 –> 0:58:44.54 - I think there's another question in the chat.

0:58:44.54 –> 0:58:46.41 - Yeah, so this is a great question.

0:58:46.41 –> 0:58:48.4 I'm glad shows you on.

0:58:48.4 –> 0:58:50.22 I hope I got that.

0:58:50.22 –> 0:58:52.53 It seems there are multiple ways to calculate the Tate

0:58:52.53 –> 0:58:55.42 from standardization to waiting to the outcome model.

0:58:55.42 –> 0:58:57.42 Do you have comments for their performance under different

0:58:57.42 –> 0:58:58.42 circumstances?

0:58:58.42 –> 0:59:00.59 Great question, and I don't.

0:59:00.59 –> 0:59:01.89 I mean, there has been...

0:59:01.89 –> 0:59:03.9 This is an area where I think

0:59:03.9 –> 0:59:06.3 it'd be great to have more research on this topic.

0:59:06.3 –> 0:59:09.49 So I have this one paper with Holger Kern and Jennifer Hill

0:59:09.49 –> 0:59:14.08 where we sort of did try to kind of explore that.

0:59:14.08 –> 0:59:16.09 And honestly, what we found not surprisingly

0:59:16.09 –> 0:59:20.08 is that if that no unmeasured moderator assumption holds,

0:59:20.08 –> 0:59:22.65 all the different methods are pretty good and fine.

0:59:22.65 –> 0:59:25.03 And like, we didn't see much difference in them.

0:59:25.03 –> 0:59:27.65 If that no unobserved moderator assumption doesn't hold

0:59:27.65 –> 0:59:28.84 then of course, none of them are good.

0:59:28.84 –> 0:59:31.843 So it sort of is like similar to propensity score world.

0:59:33.097 –> 0:59:35.24 Like, the data you have is more important than what you do

0:59:35.24 –> 0:59:36.653 with the data in a sense.

0:59:37.54 –> 0:59:39.73 But anyway, I think that that is something that like,

0:59:39.73 –> 0:59:41.535 we need a lot more work on.

0:59:41.535 –> 0:59:44.64 One thing, for example, I do have a student working on this.

0:59:44.64 –> 0:59:47.48 Like, we're trying to see if your sample

0:59:47.48 –> 0:59:50.63 is a tiny proportion of the population, like how. . .

0:59:50.63 –> 0:59:51.67 Cause like there's different.

0:59:51.67 –> 0:59:54.25 That's one where like waiting might not work as well

0:59:54.25 –> 0:59:55.25 actually, who knows.

0:59:56.26 –> 0:59:58.32 Anyways, so like all of these different data scenarios,

0:59:58.32 –> 1:00:00.86 I think need a lot more investigation to have better

1:00:00.86 –> 1:00:03.743 guidance on when the different methods work well.

1:00:09.39 –> 1:00:10.95 Anything else or maybe we're out of time?

1:00:10.95 –> 1:00:13.953 I don't know, how tight you are at one o'clock.

1:00:20.03 –> 1:00:21.98 - I think we're at an hour, so let's. . .