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 --> 00:00:36.800 on this slide.

00:00:36.800 --> 00:00:38.890 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, 00:01:06.170 --> 00:01:08.310 but there hasn't been a lot of work thinking about it

00:01:08.310 --> 00:01:09.143 quantitatively.

00:01:09.143 --> 00:01:11.380 So again, my goal today will be to sort of help

00:01:11.380 --> 00:01:14.840 give a framework for thinking about external validity

00:01:14.840 --> 00:01:16.863 in sort of a more formal way.

00:01:18.900 --> 00:01:22.380 So let's start out with the sorts of questions

00:01:22.380 --> 00:01:25.220 that might be relevant when you're thinking about

00:01:25.220 --> 00:01:26.740 external validity.

00:01:26.740 --> 00:01:30.080 So it might be research questions like a health insurer

00:01:30.080 --> 00:01:33.720 is deciding whether or not to approve some new treatment

00:01:33.720 --> 00:01:35.890 for back pain.

00:01:35.890 --> 00:01:39.090 There might be interested predicting overall population

00:01:39.090 --> 00:01:43.140 impacts of a broad public health media campaign.

00:01:43.140 --> 00:01:46.130 A physician practice might be deciding whether training

00:01:46.130 --> 00:01:48.770 providers in a new intervention would actually be cost

00:01:48.770 --> 00:01:52.630 effective given the patient population that they have.

00:01:52.630 --> 00:01:54.970 And that I felt like I needed to get some COVID

00:01:54.970 --> 00:01:57.300 example in...

00:01:57.300 --> 00:01:59.290 But, for example, a healthcare system,

00:01:59.290 --> 00:02:02.080 might wanna know whether it's sort of giving convalescent

00:02:02.080 --> 00:02:05.690 plasma to all of the individuals recently diagnosed

00:02:05.690 --> 00:02:08.240 with COVID-19 in their system, whether that would

00:02:08.240 --> 00:02:10.853 sort of lead to better outcomes overall.

00:02:12.470 --> 00:02:14.560 So all of these...

00:02:14.560 --> 00:02:16.860 What I'm distinguishing here or sort of trying to convey

00:02:16.860 --> 00:02:19.880 is that all of these reflect what I will call a population

00:02:19.880 --> 00:02:21.500 average treatment effect.

00:02:21.500 --> 00:02:24.640 So across some well-defined population,

00:02:24.640 --> 00:02:28.240 does some intervention work sort of on average.

00:02:28.240 --> 00:02:30.210 The population might be pretty narrow.

00:02:30.210 --> 00:02:33.130 Again, it might be the patients in one particular

00:02:33.130 --> 00:02:35.490 physician practice, or might be quite broad.

00:02:35.490 --> 00:02:38.140 It could be everyone in the State of Connecticut

00:02:38.140 --> 00:02:40.390 or in the entire country.

00:02:40.390 --> 00:02:44.240 But either way, it's a well-defined kind of population

00:02:44.240 --> 00:02:46.080 and we'll come back to that.

00:02:46.080 --> 00:02:47.500 What's really important,

00:02:47.500 --> 00:02:50.020 and this will sort of underlie much of the talk

00:02:50.020 --> 00:02:52.480 is that kind of the whole point is that there might

00:02:52.480 --> 00:02:54.610 be underlying treatment effect heterogeneity.

00:02:54.610 --> 00:02:56.890 So there might be some individuals

00:02:56.890 --> 00:02:59.100 for whom this treatment of interest is actually

00:02:59.100 --> 00:03:01.070 more effective than others.

00:03:01.070 --> 00:03:04.410 But what I wanna be clear about, is the goal of inference

00:03:04.410 --> 00:03:06.980 that I'm talking about today, is gonna be about

00:03:06.980 --> 00:03:08.750 this overall population average.

00:03:08.750 --> 00:03:11.450 So we're not trying to say like which people

00:03:11.450 --> 00:03:14.410 are gonna benefit more or sort of to which people

00:03:14.410 --> 00:03:15.970 should we give this treatment.

00:03:15.970 --> 00:03:19.560 It's really more a question of sort of more population

00:03:19.560 --> 00:03:21.530 level decisions, sort of if we have...

00:03:21.530 --> 00:03:23.650 If we're making a decision, that's sort of a policy

00:03:23.650 --> 00:03:25.250 kind of population level,

00:03:25.250 --> 00:03:28.350 on average is this gonna be something that makes sense.

00:03:28.350 --> 00:03:30.420 So I hope that distinction makes sense.

00:03:30.420 --> 00:03:32.343 I'm happy to come back to that.

00:03:35.360 --> 00:03:38.243 So again until I don't know, five or,

00:03:38.243 --> 00:03:41.090 well maybe now more than 10 years ago,

00:03:41.090 --> 00:03:42.990 there had been relatively little attention

00:03:42.990 --> 00:03:46.470 to the question of how well results from

00:03:46.470 --> 00:03:50.040 kind of well-designed studies like a randomized trial

00:03:50.040 --> 00:03:52.920 might carry over to a relevant target population.

00:03:52.920 --> 00:03:55.830 I think in much of statistics as well as fields

00:03:55.830 --> 00:04:00.120 like education research, public policy, even healthcare,

00:04:00.120 --> 00:04:02.560 there's really been a focus on randomized trials

00:04:02.560 --> 00:04:04.950 and getting internal validity,

00:04:04.950 --> 00:04:07.440 and I'll formalize this in a minute.

00:04:07.440 --> 00:04:09.930 But in the past 10 or so years, there's been more and more

00:04:09.930 --> 00:04:13.180 interest in this idea of how well can we take the results

00:04:13.180 --> 00:04:17.030 from a particular study and then project them

00:04:17.030 --> 00:04:19.620 to well-defined target population.

00:04:19.620 --> 00:04:21.330 And again, so today I'm gonna try to give

00:04:21.330 --> 00:04:24.100 sort of an overview of the thinking in this area,

00:04:24.100 --> 00:04:26.930 along with some of the limitations and in particular,

00:04:26.930 --> 00:04:29.780 the data limitations that we have in thinking about this.

00:04:32.840 --> 00:04:35.720 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 implement

00:05:02.530 --> 00:05:05.880 this, the problem is that you need a lot more teachers

00:05:05.880 --> 00:05:08.040 to kind of roll that out statewide.

00:05:08.040 --> 00:05:10.720 And so, it led actually to a different pool of teachers

00:05:10.720 --> 00:05:11.553 being hired.

00:05:11.553 --> 00:05:13.970 And so, there's sort of scale-up problems

00:05:13.970 --> 00:05:16.170 sometimes with the interventions and that might lead

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 differences

00:05:21.250 --> 00:05:23.503 between a sample and a population.

00:05:24.770 --> 00:05:27.630 Their difference is in sort of baseline characteristics,

00:05:27.630 --> 00:05:28.757 that moderate treatment effects.

00:05:28.757 --> 00:05:31.763 And again, I'll formalize this a little bit as we go along.

00:05:32.830 --> 00:05:34.230 Just as a little bit of an aside,

00:05:34.230 --> 00:05:36.830 but in case some of you know this field a little bit,

00:05:36.830 --> 00:05:38.740 just to give you a little, just...

00:05:38.740 --> 00:05:40.000 I wanna flag this.

00:05:40.000 --> 00:05:42.810 Some people might use the term transportability.

00:05:42.810 --> 00:05:45.720 So some of the literature in this field uses the term

00:05:45.720 --> 00:05:47.170 transportability.

00:05:47.170 --> 00:05:50.090 I tend to use generalizability.

00:05:50.090 --> 00:05:51.920 There's some subtle differences between the two,

00:05:51.920 --> 00:05:55.460 which we can come back to, but for all intents and purposes,

00:05:55.460 --> 00:05:58.660 like they basically can think of them interchangeably

00:05:58.660 --> 00:06:00.210 for now.

00:06:00.210 --> 00:06:02.050 I also wanna note, if any of you kind of come

00:06:02.050 --> 00:06:05.930 from like a survey world, these debates about

00:06:05.930 --> 00:06:09.330 kind of how well a particular sample reflects a target

00:06:09.330 --> 00:06:12.200 population are exactly, not exactly the same,

00:06:12.200 --> 00:06:14.950 but very similar to the debates happening in the survey

00:06:14.950 --> 00:06:18.850 world around non-probability samples and sort of concerns

00:06:18.850 --> 00:06:19.683 about,

00:06:20.850 --> 00:06:24.760 the use of like say online surveys and things that might not

00:06:24.760 --> 00:06:28.350 have a true formal sort of survey sampling design,

00:06:28.350 --> 00:06:30.810 and sort of some of the concerns that arise about

00:06:30.810 --> 00:06:31.643 generalizability.

00:06:31.643 --> 00:06:34.110 So there's this whole parallel literature in the survey

00:06:34.110 --> 00:06:34.990 world.

00:06:34.990 --> 00:06:36.950 Andrew Mercer has a nice summary of that.

00:06:36.950 --> 00:06:39.123 Again, I'm happy to talk more about that.

00:06:41.390 --> 00:06:43.803 Okay, any questions before I keep going?

00:06:48.500 --> 00:06:49.440 Okay.

00:06:49.440 --> 00:06:52.350 So let me formalize kind of what we're talking about

00:06:52.350 --> 00:06:53.480 a little bit.

00:06:53.480 --> 00:06:54.660 This is...

00:06:54.660 --> 00:06:59.200 This framework is now, 12 years old.

00:06:59.200 --> 00:07:00.550 Time goes quickly.

00:07:00.550 --> 00:07:04.660 But we're just to formalize what we're interested in.

00:07:04.660 --> 00:07:07.090 The goal is to estimate, again, this what I'll call

00:07:07.090 --> 00:07:09.483 a population average treatment effect or PATE.

00:07:10.440 --> 00:07:12.000 And so here,

00:07:12.000 --> 00:07:14.360 hopefully you're familiar with sort of potential outcomes

00:07:14.360 --> 00:07:15.910 and causal inference.

00:07:15.910 --> 00:07:18.780 But the idea is that we have some well-defined population

00:07:18.780 --> 00:07:20.100 of size N.

00:07:20.100 --> 00:07:23.760 And Y(1) is the potential outcomes, if people

00:07:23.760 --> 00:07:27.790 in that population receive the treatment condition

00:07:27.790 --> 00:07:29.050 of interest.

00:07:29.050 --> 00:07:31.860 Y(0) are the outcomes if they receive the control

00:07:31.860 --> 00:07:33.890 or comparison condition of interest.

00:07:33.890 --> 00:07:35.400 So here, we're just saying we're interested

00:07:35.400 --> 00:07:39.750 in the average effect, basically sort of the difference

00:07:39.750 --> 00:07:44.463 in potential outcomes, average across the population.

00:07:45.530 --> 00:07:49.330 We could be doing this with risk ratios

00:07:49.330 --> 00:07:51.450 or odds ratios or something.

00:07:51.450 --> 00:07:53.150 Those are a little more complicated because the math

00:07:53.150 --> 00:07:55.120 doesn't work as nicely.

00:07:55.120 --> 00:07:57.380 So for now think about it more like risk differences

00:07:57.380 --> 00:07:59.500 or something, if you have a binary outcome, 00:07:59.500 --> 00:08:01.573 the same fundamental points hold.

00:08:02.570 --> 00:08:05.070 So I'm not gonna tell you right now where

00:08:05.070 --> 00:08:08.010 the data we have came from, but imagine that we just

00:08:08.010 --> 00:08:10.510 have a simple estimate of this PATE,

00:08:10.510 --> 00:08:13.670 as the difference in means of some outcome

00:08:13.670 --> 00:08:16.180 between an observed treated group and an observed

00:08:16.180 --> 00:08:17.180 control group.

00:08:17.180 --> 00:08:19.520 So again, we see that there's a bunch of people

00:08:19.520 --> 00:08:22.010 who got treated, a bunch of people who got control,

00:08:22.010 --> 00:08:25.350 and we might estimate this PATE as just the simple

00:08:25.350 --> 00:08:27.850 difference in means between again, the treatment group

00:08:27.850 --> 00:08:29.350 and the control group.

00:08:29.350 --> 00:08:31.560 So what I wanna talk through for the next couple of minutes,

00:08:31.560 --> 00:08:35.930 is the bias in this sort of naive estimate of the PATE.

00:08:35.930 --> 00:08:37.940 So we'll call that Delta.

00:08:37.940 --> 00:08:40.150 So I'm being a little loose with notation here,

00:08:40.150 --> 00:08:43.270 but sort of the PATE that the bias essentially

00:08:43.270 --> 00:08:45.170 think of it as sort of the difference between

00:08:45.170 --> 00:08:49.240 the true population effect and our naive estimate of it.

00:08:49.240 --> 00:08:53.950 And what this paper did with Gary King and Kosuke Imai,

00:08:53.950 --> 00:08:58.380 we sort of laid how different choices of study designs

00:08:58.380 --> 00:09:00.840 impact the size of this bias.

00:09:00.840 --> 00:09:02.610 And in particular, we showed that sort of under

00:09:02.610 --> 00:09:05.470 some simplifying situations, 00:09:05.470 --> 00:09:07.400 sort of mathematical simplicity,

00:09:07.400 --> 00:09:11.080 you can decompose that overall bias into four pieces.

00:09:11.080 --> 00:09:15.360 So the two Delta S terms are what are called,

00:09:15.360 --> 00:09:17.450 what we call sample selection bias.

00:09:17.450 --> 00:09:22.090 So basically, the bias that comes in if our data sample

00:09:22.090 --> 00:09:24.790 is not representative of the target population

00:09:24.790 --> 00:09:25.740 that we care about.

00:09:26.750 --> 00:09:31.300 The Delta T terms are our typical sort of confounding bias.

00:09:31.300 --> 00:09:35.670 So bias that comes in if our treatment group is dissimilar

00:09:35.670 --> 00:09:36.863 from our control group.

00:09:37.870 --> 00:09:40.340 The X refers to the variables we observe,

00:09:40.340 --> 00:09:43.373 and the U refers to variables that we don't observe.

00:09:44.670 --> 00:09:46.280 So what we then did in the paper,

00:09:46.280 --> 00:09:49.220 and this is sort of what motivates a lot of this work

00:09:49.220 --> 00:09:51.370 is to think through these, again, the trade offs

00:09:51.370 --> 00:09:53.200 in these different designs.

00:09:53.200 --> 00:09:56.080 And essentially what we're trying to sort of point out

00:09:56.080 --> 00:09:57.160 is that...

00:09:58.860 --> 00:10:01.190 Let's go to the second row of this table first actually,

00:10:01.190 --> 00:10:02.460 a typical experiment.

00:10:02.460 --> 00:10:05.600 So a typical experiment, I would say is one where

00:10:05.600 --> 00:10:08.050 we kind of take whoever comes in the door,

00:10:08.050 --> 00:10:11.220 we kind of try to recruit people for a randomized trial,

00:10:11.220 --> 00:10:16.220 whether that's schools or patients or whatever it is.

00:10:16.420 --> 00:10:18.810 And we randomized them to treatment and control groups.

00:10:18.810 --> 00:10:21.060 So that is our typical randomized experiment.

00:10:22.100 --> 00:10:26.380 The treatment selection bias in that case is zero.

00:10:26.380 --> 00:10:29.140 In expectation, that's why we like randomized experiments.

00:10:29.140 --> 00:10:31.810 In expectation, there is no confounding

00:10:31.810 --> 00:10:34.300 and we get an unbiased treatment effect estimate

00:10:34.300 --> 00:10:36.670 for the sample at hand.

00:10:36.670 --> 00:10:39.830 The problem for population inference

00:10:39.830 --> 00:10:43.300 is that the Delta S terms might be big,

00:10:43.300 --> 00:10:46.230 because the people that agree to be in a randomized trial,

00:10:46.230 --> 00:10:49.100 might be quite different from the overall population

00:10:49.100 --> 00:10:50.630 that we care about.

00:10:50.630 --> 00:10:53.010 So in this paper, we're trying to just sort of...

00:10:53.010 --> 00:10:55.650 In some ways, be a little provocative and point this out

00:10:55.650 --> 00:10:59.430 that our standard thinking about study designs

00:10:59.430 --> 00:11:03.240 and sort of our prioritization of randomized trials,

00:11:03.240 --> 00:11:07.130 implicitly prioritizes internal validity over external

00:11:07.130 --> 00:11:08.400 validity.

00:11:08.400 --> 00:11:12.030 And in particular, if we really care about

00:11:12.030 --> 00:11:15.010 population effects, we really should be thinking about

00:11:15.010 --> 00:11:18.200 these together and trying to sort of have small

00:11:18.200 --> 00:11:21.820 sample selection bias and small treatment selection bias.

00:11:21.820 --> 00:11:25.450 So an ideal experiment would be one where we can randomly

00:11:25.450 --> 00:11:27.610 select people for our trial.

00:11:27.610 --> 00:11:29.840 Let's say we have...

00:11:29.840 --> 00:11:31.060 Well, actually, I'll come back to that in a second.

00:11:31.060 --> 00:11:34.020 Randomly select people for our trial and then randomly

00:11:34.020 --> 00:11:36.560 assign people to treatment or control groups.

00:11:36.560 --> 00:11:40.680 And in expectation, we will have zero bias in our population

00:11:40.680 --> 00:11:42.240 effect estimate.

00:11:42.240 --> 00:11:43.970 But these other designs, and again,

00:11:43.970 --> 00:11:47.040 like a typical experiment might end up having larger bias

00:11:47.040 --> 00:11:50.910 overall, than a well designed non-experimental study,

00:11:50.910 --> 00:11:53.650 where if we do a really good job like adjusting

00:11:53.650 --> 00:11:55.250 for confounders,

00:11:55.250 --> 00:11:59.270 it may be that well done non-experimental study

00:11:59.270 --> 00:12:01.940 conducted using say the electronic health records

00:12:01.940 --> 00:12:05.700 from a healthcare system might actually give us lower bias

00:12:05.700 --> 00:12:08.290 for a population effect estimate.

00:12:08.290 --> 00:12:12.120 Then does a non-representative small randomized trial.

00:12:12.120 --> 00:12:13.480 Again, a little provocative,

00:12:13.480 --> 00:12:16.670 but I think useful to be thinking about what is really our

00:12:16.670 --> 00:12:19.340 target of inference and how do we get data that is most

00:12:19.340 --> 00:12:20.513 relevant for that.

00:12:21.570 --> 00:12:24.260 I will also just as a small aside,

00:12:24.260 --> 00:12:25.740 maybe a little on the personal side,

00:12:25.740 --> 00:12:28.430 but it's been striking to me in the past two days.

00:12:28.430 --> 00:12:31.300 So my husband broke his collarbone over the weekend.

00:12:31.300 --> 00:12:34.730 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 --> 00:12:58.750 Oh, no, I didn't come home, we were home all the time.

00:12:58.750 --> 00:13:00.320 I asked my husband later, I was like,

00:13:00.320 --> 00:13:02.380 would you ever agree to be randomized?

00:13:02.380 --> 00:13:04.720 Like right now, we are trying to make this decision about,

00:13:04.720 --> 00:13:06.770 should you have surgery or not.

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

00:13:33.410 --> 00:13:36.893 of that when we do wanna make population inferences.

00:13:38.230 --> 00:13:39.063 Make sense so far?

00:13:39.063 --> 00:13:41.253 I can't see faces, so hopefully.

00:13:43.290 --> 00:13:44.123 Okay.

00:13:46.500 --> 00:13:47.580 So,

00:13:47.580 --> 00:13:50.270 I will say a lot of my work in this area has actually,

00:13:50.270 --> 00:13:53.480 in part been just helping or trying to raise awareness

00:13:53.480 --> 00:13:55.980 of thinking about external validity bias.

00:13:55.980 --> 00:13:59.900 So some of the research in this area has been trying

00:13:59.900 --> 00:14:02.520 to understand how big of a problem is this.

00:14:02.520 --> 00:14:05.960 If maybe people don't agree to be in randomized trials

00:14:05.960 --> 00:14:07.170 very often,

00:14:07.170 --> 00:14:09.810 but maybe that doesn't really cause bias in terms

00:14:09.810 --> 00:14:12.300 of our population effect estimates.

00:14:12.300 --> 00:14:14.670 So what I've done in a couple of papers on these

00:14:14.670 --> 00:14:18.240 other sides on this slide is basically trying to formalize

00:14:18.240 --> 00:14:22.170 this and it's pretty intuitive, but basically we show,

00:14:22.170 --> 00:14:24.150 and I'm not showing you the formulas here.

00:14:24.150 --> 00:14:27.910 But intuitively, there will be bias in a population effect

00:14:27.910 --> 00:14:31.550 estimate essentially if participation in the trial

00:14:32.590 --> 00:14:35.210 is associated with the size of the impacts.

00:14:35.210 --> 00:14:36.563 So in particular,

00:14:37.510 --> 00:14:39.250 what I'll call the external validity bias.

00:14:39.250 --> 00:14:40.083 So,

00:14:40.083 --> 00:14:42.150 those Delta S terms kind of the bias

00:14:42.150 --> 00:14:44.720 due to the lack of representativeness

00:14:44.720 --> 00:14:47.520 is a function of the variation of the probabilities

00:14:47.520 --> 00:14:49.640 of participating in a trial,

00:14:49.640 --> 00:14:51.540 variation and treatment effects,

00:14:51.540 --> 00:14:54.190 and then the correlation between those things.

00:14:54.190 --> 00:14:55.770 So if constant...

00:14:55.770 --> 00:14:57.640 If we have treat constant treatment effects

00:14:57.640 --> 00:14:59.430 or the treatment effect is zero

00:14:59.430 --> 00:15:02.340 or is two for everyone, there's gonna be no external

00:15:02.340 --> 00:15:03.173 validity bias.

00:15:03.173 --> 00:15:04.960 It doesn't matter who is in our study.

00:15:06.300 --> 00:15:07.520 Or if there...

00:15:07.520 --> 00:15:10.030 If everyone has an equal probability of participating

00:15:10.030 --> 00:15:13.770 in the study, we really do have a nice random selection,

00:15:13.770 --> 00:15:17.120 then again, there's gonna be no external validity bias.

00:15:17.120 --> 00:15:19.890 Or if the factors that influence whether or not you

00:15:19.890 --> 00:15:23.440 participate in the study are independent of the factors

00:15:23.440 --> 00:15:25.150 that moderate treatment effects,

00:15:25.150 --> 00:15:27.803 again, there'll be no external validity bias.

00:15:28.810 --> 00:15:32.250 The problem is that we often have very limited information

00:15:32.250 --> 00:15:33.920 about these pieces.

00:15:33.920 --> 00:15:37.940 We, as a field, I think medicine, public health, education,

00:15:37.940 --> 00:15:41.010 all the fields I worked in, there has not been much

00:15:41.010 --> 00:15:44.200 attention paid to these processes of how we actually

00:15:44.200 --> 00:15:45.970 enroll people in studies.

00:15:45.970 --> 00:15:49.080 And so it's hard to know kind of what factors relate

00:15:49.080 --> 00:15:52.030 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 document this

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-experimental studies

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 population.

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 trials

00:17:11.810 --> 00:17:14.760 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 representative

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

00:17:57.740 --> 00:17:59.240 the people in the sort of what they call

00:17:59.240 --> 00:18:02.930 the efficacy sample, those sort of typical trial sample

00:18:02.930 --> 00:18:05.490 had better outcomes and larger treatment effects

00:18:05.490 --> 00:18:08.853 than the overall pragmatic trial sample as a whole.

00:18:10.340 --> 00:18:14.590 We did something similar sort of in education research where

00:18:15.450 --> 00:18:16.480 it's a little bit in the weeds.

00:18:16.480 --> 00:18:17.850 I don't really wanna get into the details,

00:18:17.850 --> 00:18:22.050 but we essentially had a pretty reasonable regression

00:18:22.050 --> 00:18:23.290 discontinuity design.

00:18:23.290 --> 00:18:26.180 So we were able to get estimates of the effects of this

00:18:26.180 --> 00:18:30.030 reading first intervention across a number of states.

00:18:30.030 --> 00:18:33.780 And we then compared those state wide impact estimates

00:18:33.780 --> 00:18:37.690 to the estimates you would get if we enrolled only

00:18:37.690 --> 00:18:40.730 the sorts of schools and school districts that are typically

00:18:40.730 --> 00:18:44.110 included in educational evaluations.

00:18:44.110 --> 00:18:47.640 And there we found that this external validity bias

00:18:47.640 --> 00:18:50.040 was about 0.1 standard deviations,

00:18:50.040 --> 00:18:52.970 which in education world is fairly large.

00:18:52.970 --> 00:18:55.660 Certainly people would be concerned about an internal

00:18:55.660 --> 00:18:57.530 validity bias of that size.

00:18:57.530 --> 00:18:59.710 So we were able to sort of use this to say, look,

00:18:59.710 --> 00:19:03.010 if we really wanna be serious about external validity, 00:19:03.010 --> 00:19:06.400 it might be as much of a problem as sort of typical internal

00:19:06.400 --> 00:19:09.353 validity bias that people care about in that field.

00:19:12.740 --> 00:19:14.530 So again, the problem though, is we don't usually

00:19:14.530 --> 00:19:16.900 have these sorts of designs where we have a population

00:19:16.900 --> 00:19:18.990 effect estimate, and then sample estimates,

00:19:18.990 --> 00:19:20.620 and we can compare them.

00:19:20.620 --> 00:19:23.860 And so instead we can sometimes try to get evidence on sort

00:19:23.860 --> 00:19:24.693 of the pieces.

00:19:24.693 --> 00:19:27.630 So, but again, we basically often have very little

00:19:27.630 --> 00:19:31.350 information on why people end up participating in trials.

00:19:31.350 --> 00:19:33.730 And we also are having,

00:19:33.730 --> 00:19:36.260 I think there's growing numbers of methods,

00:19:36.260 --> 00:19:38.570 but there's still limited information on treatment effect

00:19:38.570 --> 00:19:40.010 heterogeneity.

00:19:40.010 --> 00:19:42.570 Individual randomized trials are almost never powered

00:19:42.570 --> 00:19:45.240 to detect subgroup effects.

00:19:45.240 --> 00:19:47.760 Although, there is really growing research in this field

00:19:47.760 --> 00:19:50.193 and that is maybe a topic for another day.

00:19:52.380 --> 00:19:53.400 Okay.

00:19:53.400 --> 00:19:54.980 But again, there is a little...

00:19:54.980 --> 00:19:57.900 I think I'll go through this really quickly, but,

00:19:57.900 --> 00:20:01.110 I will give credit to some fields which are trying to better

00:20:01.110 --> 00:20:04.010 understand kind of who are the people that enroll in trials

00:20:04.010 --> 00:20:08.030 and how do they compare policy populations of interest.

00:20:08.030 --> 00:20:10.620 So a lot of that has been done in sort of the substance

00:20:10.620 --> 00:20:11.710 use field.

00:20:11.710 --> 00:20:14.240 And you can see a bunch of sites here

00:20:14.240 --> 00:20:17.970 documenting that people who participate in randomized trials

00:20:17.970 --> 00:20:21.760 of substance use treatment do actually differ quite

00:20:21.760 --> 00:20:25.050 substantially from people seeking treatment for substance

00:20:25.050 --> 00:20:26.880 use problems more generally.

00:20:26.880 --> 00:20:31.640 So for example, the Okuda reference the eligibility criteria

00:20:31.640 --> 00:20:35.510 in cannabis treatment RCTs would exclude about 80%

00:20:35.510 --> 00:20:38.160 of patients across the U.S. seeking treatment

00:20:38.160 --> 00:20:39.960 for cannabis use.

00:20:39.960 --> 00:20:42.900 And so again, it's sort of there's indications

00:20:42.900 --> 00:20:45.220 that the people that participate in trials

00:20:45.220 --> 00:20:47.900 are not necessarily reflective of the people

00:20:47.900 --> 00:20:50.183 for whom decisions are having to be made.

00:20:53.920 --> 00:20:57.420 Okay, so hopefully that at least kind of give some

00:20:57.420 --> 00:21:00.740 motivation for why we want to think more carefully

00:21:00.740 --> 00:21:03.630 about the population average treatment effect

00:21:03.630 --> 00:21:05.920 and why we might wanna think about designing studies

00:21:05.920 --> 00:21:09.670 or analyzing data in ways that help us estimate that.

00:21:09.670 --> 00:21:12.683 Any questions before I move to, how do we do that?

00:21:18.590 --> 00:21:19.910 Okay.

00:21:19.910 --> 00:21:21.090 I will end...

00:21:21.090 --> 00:21:24.370 I'm gonna hopefully end it at about 12:45, 1250,

00:21:24.370 --> 00:21:26.043 so we'll have time at the end, too.

00:21:27.461 --> 00:21:30.840 So, as a statistician, I feel obligated to say,

00:21:30.840 --> 00:21:32.270 and actually I have a quote on this at the very end

00:21:32.270 --> 00:21:33.420 of the talk.

00:21:33.420 --> 00:21:35.780 If we wanna be serious about estimating something,

00:21:35.780 --> 00:21:38.460 it's better to incorporate that through the design

00:21:38.460 --> 00:21:41.110 of our study, rather than trying to do it post talk

00:21:41.110 --> 00:21:41.943 at the end.

00:21:43.670 --> 00:21:46.730 So let's talk briefly about how we can improve external

00:21:46.730 --> 00:21:49.933 validity through study or randomized trial design.

00:21:51.687 --> 00:21:52.690 So again,

00:21:52.690 --> 00:21:55.990 as I alluded to earlier with the sort of ideal experiment.

00:21:55.990 --> 00:21:59.210 An ideal scenario is one where we can randomly sample

00:21:59.210 --> 00:22:02.480 from a population and then randomly assign treatment

00:22:02.480 --> 00:22:04.070 and control conditions.

00:22:04.070 --> 00:22:07.430 Doing this will give us a formerly unbiased treatment effect

00:22:07.430 --> 00:22:10.080 estimate in the population of interest.

00:22:10.080 --> 00:22:11.240 This is wonderful.

00:22:11.240 --> 00:22:14.703 I know of about six examples of this type.

00:22:16.960 --> 00:22:19.310 Most of the examples I know of are actually a federal

00:22:19.310 --> 00:22:22.660 government programs where they are administered through

00:22:22.660 --> 00:22:24.670 like centers or sites.

00:22:24.670 --> 00:22:27.960 And the federal government was able to mandate participation

00:22:27.960 --> 00:22:29.140 in an evaluation.

00:22:29.140 --> 00:22:32.750 So classic example is the Head Start Impact Study, 00:22:32.750 --> 00:22:36.420 where they were able to randomly select headstart centers

00:22:36.420 --> 00:22:37.260 to participate.

00:22:37.260 --> 00:22:39.260 And then within each center,

00:22:39.260 --> 00:22:42.290 they randomized kids to be able to get in off the wait list

00:22:42.290 --> 00:22:43.760 versus not.

00:22:43.760 --> 00:22:46.763 An upward bound evaluation had a very similar design.

00:22:47.730 --> 00:22:49.780 It's funny, I was...

00:22:49.780 --> 00:22:52.360 I gave a talk on this topic at Facebook and I was like,

00:22:52.360 --> 00:22:54.210 why is Facebook gonna care about this?

00:22:54.210 --> 00:22:56.100 Because you would think at a place like Facebook,

00:22:56.100 --> 00:22:58.540 they have their user sample,

00:22:58.540 --> 00:23:01.850 they should be able to do randomization within,

00:23:01.850 --> 00:23:04.180 like they should be able to pick users randomly

00:23:04.180 --> 00:23:06.360 and then do any sort of random assignment they want

00:23:06.360 --> 00:23:07.200 within that.

00:23:07.200 --> 00:23:10.270 It turns out it's more complicated than that, and so,

00:23:10.270 --> 00:23:12.000 they were interested in this topic,

00:23:12.000 --> 00:23:14.590 but I think that's another sort of example where people

00:23:14.590 --> 00:23:16.490 should be thinking, could we do this?

00:23:16.490 --> 00:23:17.520 Like,

00:23:17.520 --> 00:23:18.653 in a health system.

00:23:19.640 --> 00:23:22.390 I can imagine Geisinger or something implement something

00:23:22.390 --> 00:23:24.190 in their electronic health record where

00:23:24.190 --> 00:23:25.860 it's about messaging or something.

00:23:25.860 --> 00:23:29.020 And you could imagine actually picking people randomly

00:23:29.020 --> 00:23:30.600 to then randomize.

00:23:30.600 --> 00:23:32.100 But again, that's pretty rare.

00:23:33.140 --> 00:23:35.390 There's an idea that's called purpose of sampling.

00:23:35.390 --> 00:23:39.197 And this goes back to like the 1960s or 70s

00:23:39.197 --> 00:23:43.800 and the idea is sort of picking subjects purposefully.

00:23:43.800 --> 00:23:47.210 So one example here is like maybe we think

00:23:47.210 --> 00:23:49.330 that this intervention might look different

00:23:49.330 --> 00:23:51.760 or have different effects for large versus small

00:23:51.760 --> 00:23:52.593 school districts.

00:23:52.593 --> 00:23:55.750 So in our study, we just make an effort to enroll

00:23:55.750 --> 00:23:57.803 both large and small districts.

00:23:58.720 --> 00:23:59.630 This is sort of nice.

00:23:59.630 --> 00:24:04.373 It kind of gives you some variability in the types of people

00:24:05.410 --> 00:24:08.870 or subjects in the trial, but, it doesn't have the formal

00:24:08.870 --> 00:24:11.570 representativeness and sort of the formal unbiasness,

00:24:11.570 --> 00:24:14.510 like the random sampling I just talked about.

00:24:14.510 --> 00:24:17.210 And then again, sort of similar is this idea and this push

00:24:17.210 --> 00:24:20.060 in many fields towards pragmatic or practical clinical

00:24:20.060 --> 00:24:23.610 trials, where the idea is just to sort of try to enroll

00:24:23.610 --> 00:24:26.610 like kind of more representative sample

00:24:26.610 --> 00:24:28.780 in sort of a hand wavy way like I'm doing now.

00:24:28.780 --> 00:24:31.440 So not, it doesn't have this sort of formal statistical

00:24:31.440 --> 00:24:34.640 underpinning, but at least it's trying to make sure

00:24:34.640 --> 00:24:38.020 that it's not just patients from the Yale hospital

00:24:38.020 --> 00:24:41.120 and the Hopkins hospital and whatever sort of large medical

00:24:41.120 --> 00:24:44.510 centers, at least they might be trying to enroll patients

00:24:44.510 --> 00:24:46.703 from a broader spectrum across the U.S.

00:24:48.800 --> 00:24:52.970 Unfortunately, though, as much as I want to do things

00:24:52.970 --> 00:24:55.660 for design often, we're in a case where there's a study

00:24:55.660 --> 00:25:00.110 that's already been conducted and we are just

00:25:00.110 --> 00:25:01.310 sort of stuck analyzing it.

00:25:01.310 --> 00:25:04.420 And we wanna get a sense for how representative

00:25:04.420 --> 00:25:06.893 the results might be for a population.

00:25:08.740 --> 00:25:10.340 Sometimes people, when I talk about this,

00:25:10.340 --> 00:25:12.510 people are like, well, isn't this what meta-analysis does?

00:25:12.510 --> 00:25:16.080 Like meta-analysis enables you to combine multiple

00:25:16.080 --> 00:25:19.820 randomized trials and come up with sort of an overall

00:25:19.820 --> 00:25:20.723 effect estimate.

00:25:22.650 --> 00:25:26.410 And my answer to that is sort of yes maybe, or no maybe.

00:25:26.410 --> 00:25:29.650 Basically, the challenge with meta-analysis,

00:25:29.650 --> 00:25:33.760 is that until recently, no one really had a potential target

00:25:33.760 --> 00:25:35.270 population.

00:25:35.270 --> 00:25:38.000 It was not very formal about what the target population is.

00:25:38.000 --> 00:25:41.230 I think underlying that analysis is generally

00:25:41.230 --> 00:25:43.790 sort of a belief that the effects are constant

00:25:43.790 --> 00:25:45.793 and we're just trying to pool data.

00:25:47.538 --> 00:25:48.371 And it...

00:25:48.371 --> 00:25:49.760 And even just like, you can sort of see this,

00:25:49.760 --> 00:25:52.170 like if all of the trials sampled the same

00:25:52.170 --> 00:25:54.420 non-representative population, 00:25:54.420 --> 00:25:56.980 combining them is not going to help you get towards

00:25:56.980 --> 00:25:58.143 representativeness.

00:25:59.120 --> 00:26:01.410 That's that I have a former Postdoc Hwanhee Hong,

00:26:01.410 --> 00:26:02.850 who's now at Duke.

00:26:02.850 --> 00:26:05.540 And she has been doing some work to try to bridge

00:26:05.540 --> 00:26:07.970 these worlds and sort of really try to think through,

00:26:07.970 --> 00:26:11.590 well, how can we better use multiple trials

00:26:11.590 --> 00:26:14.233 to get to target population effects?

00:26:15.520 --> 00:26:18.340 There's another field it's called risk cross-design

00:26:18.340 --> 00:26:21.060 synthesis or research synthesis.

00:26:21.060 --> 00:26:22.000 This is sort of neat.

00:26:22.000 --> 00:26:26.170 It's one where you kind of combine randomized trial data,

00:26:26.170 --> 00:26:29.820 which might be not representative with non-experimental

00:26:29.820 --> 00:26:30.653 study data.

00:26:30.653 --> 00:26:34.320 So sort of explicitly trading off the internal and external

00:26:34.320 --> 00:26:35.930 validity.

00:26:35.930 --> 00:26:37.240 I'm not gonna get into the details,

00:26:37.240 --> 00:26:38.260 there's some references here.

00:26:38.260 --> 00:26:41.360 Ellie Kaizar at Ohio State, is one of the people

00:26:41.360 --> 00:26:43.283 that's done a lot of work on this.

00:26:45.310 --> 00:26:48.180 And part of the reason I'm not focused on this is that

00:26:48.180 --> 00:26:52.510 I work in a lot of areas like education and public health,

00:26:52.510 --> 00:26:54.050 sort of social science areas,

00:26:54.050 --> 00:26:56.180 where we often don't have multiple studies.

00:26:56.180 --> 00:27:00.470 So we often are stuck with just one study and we're trying

00:27:00.470 --> 00:27:03.970 to use that to learn about target populations.

00:27:03.970 --> 00:27:07.110 So I'm gonna briefly talk about an example

00:27:07.110 --> 00:27:11.810 where we trying to sort of do this.

00:27:11.810 --> 00:27:16.200 And basically, the fundamental idea is to re-weight

00:27:16.200 --> 00:27:19.563 the study sample to look like the target population.

00:27:20.780 --> 00:27:24.960 This idea is related to post stratification

00:27:24.960 --> 00:27:27.310 or, oh my gosh, I'm blanking now.

00:27:27.310 --> 00:27:29.423 Raking adjustments in surveys.

00:27:30.660 --> 00:27:33.490 So post stratification would be sort of at a simple level,

00:27:33.490 --> 00:27:34.740 would be something like...

00:27:34.740 --> 00:27:38.300 Well, if we know that males and females

00:27:38.300 --> 00:27:41.230 have different effects, or let's say young and old

00:27:41.230 --> 00:27:43.690 have different effects, let's estimate the effects

00:27:43.690 --> 00:27:46.153 separately for young versus old.

00:27:47.130 --> 00:27:50.860 And then re-weight those using the population proportions

00:27:50.860 --> 00:27:52.683 of sort of young versus old.

00:27:54.340 --> 00:27:57.550 That sort of stratification doesn't work if you have more

00:27:57.550 --> 00:28:02.450 than like one or two categorical effect moderators.

00:28:02.450 --> 00:28:03.283 And so,

00:28:03.283 --> 00:28:05.630 what I'm gonna show today is an approach where we use

00:28:05.630 --> 00:28:07.720 weighting, where we fit a model,

00:28:07.720 --> 00:28:10.080 predicting participation in the trial,

00:28:10.080 --> 00:28:13.100 and then weight the trial sample to look like the target

00:28:13.100 --> 00:28:14.100 population.

00:28:14.100 --> 00:28:16.960 So similar idea to things like propensity score weights

00:28:16.960 --> 00:28:20.253 or non-response adjustment weights in samples.

00:28:21.370 --> 00:28:23.150 There is a different approach,

00:28:23.150 --> 00:28:26.640 So what I'm gonna illustrate today is sort of this sample

00:28:26.640 --> 00:28:29.290 selection weighting strategy.

00:28:29.290 --> 00:28:32.070 You also can tackle this external validity

00:28:32.070 --> 00:28:34.880 by trying to model the outcome very flexibly

00:28:34.880 --> 00:28:39.013 and then project outcomes in the population.

00:28:40.450 --> 00:28:42.530 In some work I did with Jennifer Hill and others,

00:28:42.530 --> 00:28:45.520 we showed that BARTs, Bayesian Additive Regression Trees

00:28:45.520 --> 00:28:47.820 can actually work quite well for that purpose.

00:28:48.920 --> 00:28:52.580 And more recently, Issa Dahabreh at Brown has done some

00:28:52.580 --> 00:28:55.240 nice work sort of bridging these two and showing

00:28:55.240 --> 00:28:58.140 basically a doubly robust kind of idea where we can use

00:28:58.140 --> 00:29:03.140 both the sample membership model and the outcome model

00:29:03.580 --> 00:29:05.660 to have better performance.

00:29:05.660 --> 00:29:08.440 But today, I'm gonna just illustrate the weighting approach,

00:29:08.440 --> 00:29:10.700 partly because it's a really nice sort of pedagogical

00:29:10.700 --> 00:29:13.540 example and helps you kind of see what's going on

00:29:13.540 --> 00:29:14.373 in the data.

00:29:15.850 --> 00:29:18.373 Okay, any questions before I continue?

00:29:20.520 --> 00:29:21.353 Okay.

00:29:22.380 --> 00:29:25.670 So the example I'm gonna use is...

00:29:25.670 --> 00:29:28.080 There was this, I mean, some of you probably know much more

00:29:28.080 --> 00:29:32.530 about HIV treatment than I do, but the ACTG Trial,

00:29:32.530 --> 00:29:35.820 which was now quite an old trial, 00:29:35.820 --> 00:29:38.590 but it was one of the ones that basically showed that

00:29:38.590 --> 00:29:41.940 HAART therapy, highly active antiretroviral therapy

00:29:41.940 --> 00:29:46.190 was quite effective at reducing time to AIDS or death

00:29:46.190 --> 00:29:49.490 compared to standard combination therapy at the time.

00:29:49.490 --> 00:29:53.910 So it randomized about 1200 U.S. HIV positive adults

00:29:53.910 --> 00:29:56.440 to treatment versus control.

00:29:56.440 --> 00:29:59.380 And the intent to tree analysis in the trial

00:29:59.380 --> 00:30:01.460 had a hazard ratio of 0.51.

00:30:01.460 --> 00:30:05.513 So again, very effective at reducing time to AIDS or death.

00:30:06.870 --> 00:30:10.400 So Steve Cole and I though kind of asked the question, well,

00:30:10.400 --> 00:30:13.010 we don't necessarily just care about the people

00:30:13.010 --> 00:30:13.920 in the trial.

00:30:13.920 --> 00:30:16.490 This seems to be a very effective treatment.

00:30:16.490 --> 00:30:19.420 What could we use this data to project out

00:30:19.420 --> 00:30:21.830 sort of what the effects of the treatment would be

00:30:21.830 --> 00:30:24.530 if it were implemented nationwide?

00:30:24.530 --> 00:30:28.400 So we from CDC got estimates of the number of people

00:30:28.400 --> 00:30:31.920 newly infected with HIV in 2006.

00:30:31.920 --> 00:30:35.230 And basically, asked the question sort of if hypothetically,

00:30:35.230 --> 00:30:39.840 everyone in that group were able to get HAART versus

00:30:39.840 --> 00:30:41.670 standard combination therapy,

00:30:41.670 --> 00:30:44.833 what would be the population impacts of this treatment?

00:30:47.700 --> 00:30:50.330 In this case, because of sort of data availability, 00:30:50.330 --> 00:30:54.630 we only had the joint distribution of age, sex and race

00:30:54.630 --> 00:30:56.070 for the population.

00:30:56.070 --> 00:30:59.370 So we made sort of a pseudo population, again,

00:30:59.370 --> 00:31:01.500 sort of representing the U.S. population

00:31:01.500 --> 00:31:03.250 of newly infected people.

00:31:03.250 --> 00:31:05.780 But again, all we have is sex, race and age,

00:31:05.780 --> 00:31:07.080 which I will come back to.

00:31:08.490 --> 00:31:11.630 So this table documents the trial and the population.

00:31:11.630 --> 00:31:14.540 So you can see for example,

00:31:14.540 --> 00:31:19.540 that the trial tended to have more sort of 30 to 39 year

00:31:19.700 --> 00:31:23.773 olds, many fewer people under 30.

00:31:24.822 --> 00:31:28.600 The trial had more males and also had more whites

00:31:28.600 --> 00:31:32.280 and fewer blacks, Hispanic was similar.

00:31:32.280 --> 00:31:35.470 But I wanna flag and we'll come back to this in a minute

00:31:35.470 --> 00:31:37.850 that, in what I'm gonna show,

00:31:37.850 --> 00:31:41.150 we can adjust for the age, sex, race distribution.

00:31:41.150 --> 00:31:43.000 But, there's a real limitation,

00:31:43.000 --> 00:31:45.960 which is that the CD4 cell count as sort of a measure

00:31:45.960 --> 00:31:50.220 of disease severity is not available in the population.

00:31:50.220 --> 00:31:53.310 So this is a potential effect moderator,

00:31:53.310 --> 00:31:56.130 which we don't observe in the population.

00:31:56.130 --> 00:31:59.340 So in sort of projecting the impacts, we can say, well,

00:31:59.340 --> 00:32:02.740 here is the predicted impact given the age, sex,

00:32:02.740 --> 00:32:05.640 race distribution, but there's this unobserved

00:32:05.640 --> 00:32:09.370 potential effect moderator that we sort of might be worried

00:32:09.370 --> 00:32:11.320 about kind of in the back of our heads.

00:32:14.560 --> 00:32:16.520 So again, I briefly mentioned this,

00:32:16.520 --> 00:32:19.750 this is like the super basic description

00:32:19.750 --> 00:32:21.780 of what can be done.

00:32:21.780 --> 00:32:24.060 There are more nuances and I have some sites at the end

00:32:24.060 --> 00:32:25.890 for sort of more details.

00:32:25.890 --> 00:32:27.780 But basically fundamentally will, again,

00:32:27.780 --> 00:32:29.700 we sort of think about it as we kind of stack

00:32:29.700 --> 00:32:30.700 our data sets together.

00:32:30.700 --> 00:32:33.750 So we put our trial sample and our population data set

00:32:33.750 --> 00:32:34.750 together.

00:32:34.750 --> 00:32:37.940 We have an indicator for whether someone is in the trial

00:32:37.940 --> 00:32:39.690 versus the population.

00:32:39.690 --> 00:32:42.530 And then, we're gonna wait the trial members

00:32:42.530 --> 00:32:45.670 by their inverse probability of being in the trial

00:32:45.670 --> 00:32:48.470 as a function of the observed covariance.

00:32:48.470 --> 00:32:51.320 And again, very similar intuition and ideas

00:32:51.320 --> 00:32:54.650 and theory underlying this as underlying things

00:32:54.650 --> 00:32:57.630 like Horvitz-Thomson estimation in sample surveys

00:32:58.480 --> 00:33:00.680 and inverse probability of treatment waiting

00:33:00.680 --> 00:33:02.363 in non-experimental studies.

00:33:06.160 --> 00:33:09.310 So I showed you earlier that age, sex and race

00:33:09.310 --> 00:33:13.320 are all related to participation in the trial.

00:33:13.320 --> 00:33:15.450 What I'm not showing you the details of,

00:33:15.450 --> 00:33:18.500 but just trust me is that those factors also moderate

00:33:18.500 --> 00:33:20.465 effects in the trial.

00:33:20.465 --> 00:33:23.960 So the trial showed the largest effects for those ages, 00:33:23.960 --> 00:33:27.620 30 to 39, males and black individuals.

00:33:27.620 --> 00:33:30.620 And so, this is exactly why then what we might think

00:33:30.620 --> 00:33:34.150 that the overall trial estimate might not reflect

00:33:34.150 --> 00:33:36.383 what we would see population-wide.

00:33:38.720 --> 00:33:40.040 Ironically though, it turns out actually

00:33:40.040 --> 00:33:41.100 it kind of all cancels out.

00:33:41.100 --> 00:33:44.910 So this table shows the estimated population effects.

00:33:44.910 --> 00:33:48.050 So the first row again, is just the sort of naive trial

00:33:48.050 --> 00:33:49.660 results.

00:33:49.660 --> 00:33:52.390 We can then sort of weight by each characteristic

00:33:52.390 --> 00:33:55.700 separately, and then the bottom row is the combined

00:33:55.700 --> 00:33:57.860 age, sex, race adjustments.

00:33:57.860 --> 00:34:00.750 And you can see sort of actually the hazard ratio

00:34:00.750 --> 00:34:02.810 was remarkably similar.

00:34:02.810 --> 00:34:04.930 It's partly because like the age weightings

00:34:04.930 --> 00:34:07.100 sort of makes the impact smaller,

00:34:07.100 --> 00:34:09.610 but then the race weighting makes it bigger.

00:34:09.610 --> 00:34:11.560 And so then it kind of just washes out.

00:34:13.270 --> 00:34:14.590 But again, it's sort of a nice example,

00:34:14.590 --> 00:34:17.010 cause you can sort of see how the patterns

00:34:17.010 --> 00:34:19.900 evolve based on the size of the effects

00:34:19.900 --> 00:34:21.423 and the sample selection.

00:34:22.550 --> 00:34:24.770 I also wanna point out though that, of course,

00:34:24.770 --> 00:34:27.470 the confidence interval is wider,

00:34:27.470 --> 00:34:30.020 and that is sort of reflecting the fact that we are doing

00:34:30.020 --> 00:34:33.260 this extrapolation from the trial sample to the population.

00:34:33.260 --> 00:34:36.210 And so there's sort of a variance price we'll pay for that.

00:34:38.990 --> 00:34:39.823 Okay.

00:34:39.823 --> 00:34:43.610 So I haven't been super formal on the assumptions,

00:34:43.610 --> 00:34:45.110 but I'm I alluded to this?

00:34:45.110 --> 00:34:47.520 So I wanna just take a few minutes to turn

00:34:47.520 --> 00:34:50.100 to what about unobserved moderators?

00:34:50.100 --> 00:34:53.770 Because again, we can interpret this 0.57

00:34:53.770 --> 00:34:58.410 as the sort of overall population effect estimate

00:34:58.410 --> 00:35:01.420 only under an assumption that there are no unobserved

00:35:01.420 --> 00:35:05.550 moderators that differ between sample and population,

00:35:05.550 --> 00:35:08.063 once we adjust for age, sex, race.

00:35:11.000 --> 00:35:12.453 Okay, and in reality,

00:35:13.500 --> 00:35:16.610 such unobserved effect moderators are likely the rule,

00:35:16.610 --> 00:35:18.340 not the exception.

00:35:18.340 --> 00:35:20.410 So again, sort of, as I just said,

00:35:20.410 --> 00:35:23.110 the key assumption is that we've basically adjusted

00:35:23.110 --> 00:35:26.460 for all of the effect moderators.

00:35:26.460 --> 00:35:29.950 Very kind of comparable assumption to the assumption

00:35:29.950 --> 00:35:33.463 of no an observed confounding in a non-experimental study.

00:35:35.040 --> 00:35:37.900 And one of the reasons this is an important assumption

00:35:37.900 --> 00:35:41.690 to think about, is that, it is quite rare actually

00:35:41.690 --> 00:35:45.570 to have extensive covariate data overlap

00:35:45.570 --> 00:35:48.070 between the sample and the population.

00:35:48.070 --> 00:35:50.650 I have been working in this area for...

00:35:50.650 --> 00:35:51.690 How many years now?

00:35:51.690 --> 00:35:52.990 At least 10 years.

00:35:52.990 --> 00:35:55.830 And I've found time and time again,

00:35:55.830 --> 00:35:58.440 across a number of content areas,

00:35:58.440 --> 00:36:01.270 that it is quite rare to have a randomized trial sample

00:36:01.270 --> 00:36:03.380 and the target population dataset

00:36:03.380 --> 00:36:06.010 with very many comparable measures.

00:36:06.010 --> 00:36:07.820 So in the Stuart and Rhodes paper,

00:36:07.820 --> 00:36:11.520 this was in like early childhood setting

00:36:11.520 --> 00:36:15.330 and each data set, the trial and the population data

00:36:15.330 --> 00:36:19.350 had like over 400 variables observed at baseline.

00:36:19.350 --> 00:36:21.990 There were literally only seven that were measured

00:36:21.990 --> 00:36:24.630 consistently between the two samples.

00:36:24.630 --> 00:36:28.120 So essentially we have very limited ability then to adjust

00:36:28.120 --> 00:36:31.403 for these factors because they just don't have much overlap.

00:36:32.290 --> 00:36:37.020 So what that then motivated us to create some sensitivity

00:36:37.020 --> 00:36:40.110 analysis to basically probe and say, well,

00:36:40.110 --> 00:36:43.230 what if there is an unobserved effect moderator,

00:36:43.230 --> 00:36:47.160 how much would that change our population effect estimate?

00:36:47.160 --> 00:36:51.370 Again, this is very comparable to analysis of sensitivity,

00:36:51.370 --> 00:36:54.350 to unobserved confounding and non-experimental studies

00:36:54.350 --> 00:36:58.680 sort of adapted for this purpose of trial population,

00:36:58.680 --> 00:36:59.683 generalized ability.

00:37:03.220 --> 00:37:05.860 I think I can skip this in the interest of time and not go

00:37:05.860 --> 00:37:06.760 through all the details.

00:37:06.760 --> 00:37:08.220 If anyone wants the slides by the way,

00:37:08.220 --> 00:37:10.520 feel free to email me, I'm happy to send them.

00:37:12.800 --> 00:37:14.720 I'm gonna skip this too cause I've already said

00:37:14.720 --> 00:37:18.780 sort of the key assumption that is relevant for right now,

00:37:18.780 --> 00:37:22.333 but basically what we propose is,

00:37:23.802 --> 00:37:25.730 I'm gonna talk about two cases.

00:37:25.730 --> 00:37:29.370 So the easier case is this one where we're gonna assume

00:37:29.370 --> 00:37:32.280 that the randomized trial observes all of the effect

00:37:32.280 --> 00:37:33.113 moderators.

00:37:33.113 --> 00:37:36.350 And the issue is that our target population dataset

00:37:36.350 --> 00:37:40.620 does not have some moderators observed.

00:37:40.620 --> 00:37:43.100 I think this is fairly realistic because at least

00:37:43.100 --> 00:37:46.590 like to think that the people running the randomized trials

00:37:46.590 --> 00:37:49.520 have enough scientific knowledge and expertise

00:37:49.520 --> 00:37:52.390 that they sort of know what the likely effect moderators

00:37:52.390 --> 00:37:54.830 are and that they measure them in the trial.

00:37:54.830 --> 00:37:57.760 That is probably not fully realistic, but I'm...

00:37:57.760 --> 00:38:00.460 I like to give them sort of the benefit of the doubt

00:38:00.460 --> 00:38:01.470 on that.

00:38:01.470 --> 00:38:04.960 And that sort of that's what the ACTG example,

00:38:04.960 --> 00:38:07.470 was like CD4 count would be an example of this,

00:38:07.470 --> 00:38:10.840 where we have CD4 count in the trial,

00:38:10.840 --> 00:38:13.520 but we just don't have it in the population.

00:38:13.520 --> 00:38:16.060 So what we showed is that there's actually,

00:38:16.060 --> 00:38:18.060 a couple of different ways you can implement

00:38:18.060 --> 00:38:20.053 this sort of sensitivity analysis.

00:38:21.510 --> 00:38:24.600 One is essentially kind of an outcome model based one

00:38:24.600 --> 00:38:25.483 where you,

00:38:27.640 --> 00:38:30.320 basically, we just sort of specify a range

00:38:30.320 --> 00:38:34.150 for the unobserved moderator V in the population.

00:38:34.150 --> 00:38:36.270 So we kind of say, well, we don't know

00:38:36.270 --> 00:38:39.780 the distribution of this moderator in the population,

00:38:39.780 --> 00:38:43.010 but we're gonna guess that it's in some range.

00:38:43.010 --> 00:38:47.860 And then, we kind of projected out using data from the trial

00:38:47.860 --> 00:38:50.540 to understand like the extent of the moderation

00:38:50.540 --> 00:38:51.743 due to that variable.

00:38:52.900 --> 00:38:55.110 There's another variation on this,

00:38:55.110 --> 00:38:57.760 which is sort of the weighting variation

00:38:57.760 --> 00:38:59.920 where you kind of adjust the weights,

00:38:59.920 --> 00:39:03.430 essentially again for this unobserved moderator.

00:39:03.430 --> 00:39:07.150 Again, either way you sort of basically just have to specify

00:39:07.150 --> 00:39:11.440 a potential range for this V, the unobserved moderator

00:39:11.440 --> 00:39:12.593 in the population.

00:39:13.960 --> 00:39:15.603 So here's an example of that.

00:39:15.603 --> 00:39:18.280 This is a different example, where we were looking

00:39:18.280 --> 00:39:21.410 at the effects of a smoking cessation intervention

00:39:21.410 --> 00:39:24.460 among people in substance use treatment.

00:39:24.460 --> 00:39:29.460 And in the randomized trial, the mean addiction score

00:39:31.300 --> 00:39:33.030 was four.

00:39:33.030 --> 00:39:34.930 But we didn't have this addiction score,

00:39:34.930 --> 00:39:37.410 in the target population of interest.

00:39:37.410 --> 00:39:40.310 And so, what the sensitivity analysis allows us to do

00:39:40.310 --> 00:39:43.760 is to say, well, let's imagine that range is anywhere

00:39:43.760 --> 00:39:45.490 from three to five.

00:39:45.490 --> 00:39:49.100 And how much does that change our population effect

00:39:49.100 --> 00:39:50.520 estimates?

00:39:50.520 --> 00:39:53.520 Essentially, how steep this line is, is gonna be

00:39:53.520 --> 00:39:56.570 sort of determine how much it matters.

00:39:56.570 --> 00:39:58.800 And the steepness of the line basically

00:39:58.800 --> 00:40:01.720 is how much of a moderator is it,

00:40:01.720 --> 00:40:05.270 sort of how much effect heterogeneity is there in the trial

00:40:05.270 --> 00:40:07.490 as a result of that variable.

00:40:07.490 --> 00:40:10.580 But again, this is at least one way to sort of turn

00:40:10.580 --> 00:40:12.970 this sort of worry about an unobserved moderator

00:40:12.970 --> 00:40:15.770 into a more formal statement about how much

00:40:15.770 --> 00:40:17.083 it really might matter.

00:40:20.946 --> 00:40:22.390 I'm not gonna get into this partly,

00:40:22.390 --> 00:40:24.300 so you might also be thinking, well,

00:40:24.300 --> 00:40:27.367 what if the trial doesn't know what all the moderators are?

00:40:27.367 --> 00:40:30.600 And what if there's some fully unobserved moderator

00:40:30.600 --> 00:40:31.773 that will call U?

00:40:33.620 --> 00:40:35.650 This is a much much harder, basically,

00:40:35.650 --> 00:40:38.688 if anyone wants to try to dig into it, that would be great.

00:40:38.688 --> 00:40:41.660 Part of the reason it's harder is because you have to make

00:40:41.660 --> 00:40:44.380 very strong assumptions about the distribution

00:40:44.380 --> 00:40:47.990 of the observed covariance and U together.

00:40:47.990 --> 00:40:49.120 We put out one approach,

00:40:49.120 --> 00:40:52.920 but it is a fairly special case and not very general.

00:40:52.920 --> 00:40:56.030 So again, hopefully we're not in this sort of scenario

00:40:56.030 --> 00:40:56.863 very often.

00:41:00.590 --> 00:41:02.560 This is a little bit of a technicality,

00:41:02.560 --> 00:41:05.330 but often epidemiologists ask this question.

00:41:05.330 --> 00:41:08.630 So I've laid stuff out again with respect to kind of a risk

00:41:08.630 --> 00:41:10.530 difference or a difference in outcomes

00:41:11.640 --> 00:41:15.090 and sort of like more of like an additive treatment scale.

00:41:15.090 --> 00:41:17.410 There is this real complication that arises,

00:41:17.410 --> 00:41:19.980 which is that if you have like a binary,

00:41:19.980 --> 00:41:24.153 like the scale of the outcome matters in terms of effect

00:41:25.160 --> 00:41:26.320 moderation.

00:41:26.320 --> 00:41:29.560 And in particular, there might be sort of more apparent

00:41:29.560 --> 00:41:32.970 effect heterogeneity on one scale versus another.

00:41:32.970 --> 00:41:36.720 So I'm just kind of flagging this, that like this exists,

00:41:36.720 --> 00:41:39.000 there are some people sort of looking at this in more

00:41:39.000 --> 00:41:44.000 formal, but again for now sort of just think about like risk

00:41:44.160 --> 00:41:45.410 difference kind of scale.

00:41:47.450 --> 00:41:48.283 Okay, great.

00:41:48.283 --> 00:41:51.400 So let me just conclude with a few kind of final thoughts.

00:41:51.400 --> 00:41:54.440 So, I think all of us, not all of us,

00:41:54.440 --> 00:41:57.610 but often we sort of want to assume that study results

00:41:57.610 --> 00:41:58.443 generalize.

00:41:58.443 --> 00:42:01.130 Often people write a discussion section in a paper,

00:42:01.130 --> 00:42:04.560 where they kind of qualitatively have some sentences

00:42:04.560 --> 00:42:07.830 about why they do or don't think that the results

00:42:07.830 --> 00:42:10.190 in this paper kind of extend to other groups

00:42:10.190 --> 00:42:11.403 or other populations.

00:42:12.520 --> 00:42:16.180 But I think until the past again, sort of five or so years,

00:42:16.180 --> 00:42:19.140 a lot of that discussion was very hand-wavy

00:42:19.140 --> 00:42:20.810 and sort of qualitative.

00:42:20.810 --> 00:42:23.540 I think that what we are seeing in epidemiology

00:42:23.540 --> 00:42:26.070 and statistics and bias statistics

00:42:26.070 --> 00:42:29.000 recently has been a push towards having more

00:42:29.000 --> 00:42:33.160 ability to quantify this and make it sort of more formal

00:42:33.160 --> 00:42:33.993 statements.

00:42:35.040 --> 00:42:37.440 So I think if we do wanna be serious though,

00:42:37.440 --> 00:42:40.590 about assessing and enhancing external validity,

00:42:40.590 --> 00:42:42.600 again, we really need these different pieces.

00:42:42.600 --> 00:42:46.040 We need information on the factors that influence effect

00:42:46.040 --> 00:42:48.540 heterogeneity the moderators.

00:42:48.540 --> 00:42:50.700 We need information on the factors that influence

00:42:50.700 --> 00:42:54.860 participation in rigorous studies like randomized trials.

00:42:54.860 --> 00:42:57.370 And we need data on all of those things,

00:42:57.370 --> 00:42:59.173 in the trial and the population.

00:43:00.380 --> 00:43:03.500 And then finally, we need statistical methods that allow us

00:43:03.500 --> 00:43:07.103 to use that data to estimate population treatment effects.

00:43:07.940 --> 00:43:11.900 I would argue that that last bullet is sort of much further

00:43:11.900 --> 00:43:13.430 along than any of the others.

00:43:13.430 --> 00:43:15.490 That in my experience,

00:43:15.490 --> 00:43:18.700 the limiting factor is usually not the methods.

00:43:18.700 --> 00:43:22.230 The limiting factor at this point in time is the data

00:43:22.230 --> 00:43:24.610 and sort of the scientific knowledge

00:43:24.610 --> 00:43:27.033 about these different factors.

00:43:29.050 --> 00:43:30.240 And that's what this slide is.

00:43:30.240 --> 00:43:32.640 So I think I've already said, but that again,

00:43:32.640 --> 00:43:35.450 is sort of one of the motivations for the sensitivity

00:43:35.450 --> 00:43:38.870 analysis is just a recognition that it's often,

00:43:38.870 --> 00:43:40.840 really quite hard to get data that

00:43:42.020 --> 00:43:45.193 is consistently measured between a trial and a population.

00:43:46.710 --> 00:43:48.730 So on that point, recommendations again,

00:43:48.730 --> 00:43:51.340 if we wanna be serious about effect heterogeneity

00:43:51.340 --> 00:43:54.780 or about estimating population treatment effects,

00:43:54.780 --> 00:43:58.170 we need better information on treatment effect heterogeneity

00:43:59.210 --> 00:44:01.690 that might be better analysis of existing trials,

00:44:01.690 --> 00:44:04.500 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.

00:45:00.630 --> 00:45:03.300 Like they standardized them to their own population,

00:45:03.300 --> 00:45:04.790 not sort of more generally.

00:45:04.790 --> 00:45:08.343 And so they, weren't sort of on the same scale in the end.

00:45:09.900 --> 00:45:12.260 As a statistician, of course, I will say we do need more

00:45:12.260 --> 00:45:15.260 research on the methods and understanding when they work

00:45:15.260 --> 00:45:16.093 and when they don't.

00:45:16.093 --> 00:45:18.630 There are some pretty strong assumptions

00:45:18.630 --> 00:45:20.350 in these approaches.

00:45:20.350 --> 00:45:23.840 But again, I think that sort of in some ways,

00:45:23.840 --> 00:45:26.893 that is further along and then some of the data situations.

00:45:28.680 --> 00:45:31.760 So I just wanted to take one minute to flag some current

00:45:31.760 --> 00:45:34.460 work in case partly if anyone wants to ask questions about

00:45:34.460 --> 00:45:36.110 these.

00:45:36.110 --> 00:45:38.220 One thing I'm kind of excited about,

00:45:38.220 --> 00:45:41.500 especially in my education world is...

00:45:41.500 --> 00:45:43.670 So what I've been talking about today has mostly been,

00:45:43.670 --> 00:45:46.010 if we have a trial sample and we wanna project

00:45:46.010 --> 00:45:48.730 to kind of a larger target population.

00:45:48.730 --> 00:45:50.710 But there's an equally interesting question,

00:45:50.710 --> 00:45:54.180 which is sort of how well can randomized trial informs

00:45:54.180 --> 00:45:55.610 or local decision making?

00:45:55.610 --> 00:46:00.043 So if we have a randomized trial with 60 schools in it,

00:46:00.990 --> 00:46:04.480 how well can the results from that trial be used to inform

00:46:04.480 --> 00:46:06.910 individual school districts decisions?

00:46:06.910 --> 00:46:08.892 Turns out, not particularly well.

00:46:08.892 --> 00:46:10.000 (laughs)

00:46:10.000 --> 00:46:11.920 We can talk more about that.

00:46:11.920 --> 00:46:15.040 I mentioned earlier, Issa Dahabreh, who's at Brown,

00:46:15.040 --> 00:46:18.100 and he's really interested in developing sort of the formal

00:46:18.100 --> 00:46:20.940 theories underlying different ways of estimating

00:46:20.940 --> 00:46:23.440 these population effects, again, including some

00:46:23.440 --> 00:46:25.163 doubly robust approaches.

00:46:26.368 --> 00:46:29.130 Trang Nguyen, who works at Hopkins with me,

00:46:29.130 --> 00:46:31.650 we are still looking at sort of the sensitivity analysis

00:46:31.650 --> 00:46:34.090 for unobserved moderators.

00:46:34.090 --> 00:46:37.190 I mentioned Hwanhee Hong already, who's now at Duke.

00:46:37.190 --> 00:46:40.450 And she, again, sort of straddles the meta-analysis world

00:46:40.450 --> 00:46:43.000 in this world, which has some really interesting

00:46:43.000 --> 00:46:43.833 connections.

00:46:44.910 --> 00:46:47.640 My former student now he's at Flatiron Health

00:46:47.640 --> 00:46:49.560 as of a few months ago.

00:46:49.560 --> 00:46:53.040 Ben Ackerman, did some work on sort of measurement error

00:46:53.040 --> 00:46:55.250 and sort of partly how to deal with some of these

00:46:55.250 --> 00:46:58.793 measurement challenges between the sample and population.

00:46:59.776 --> 00:47:03.580 And then I'll just briefly mention Daniel Westreich at UNC,

00:47:03.580 --> 00:47:05.040 who is really...

00:47:05.040 --> 00:47:08.700 If you come from sort of more of an epidemiology world,

00:47:08.700 --> 00:47:11.120 Daniel has some really nice papers that are sort of trying

00:47:11.120 --> 00:47:14.300 to translate these ideas to epidemiology,

00:47:14.300 --> 00:47:17.320 and this concept of what he calls target validity.

00:47:17.320 --> 00:47:20.250 So sort of rather than thinking about internal and external

00:47:20.250 --> 00:47:23.220 validity separately, and as potentially,

00:47:23.220 --> 00:47:25.690 in kind of conflict with each other,

00:47:25.690 --> 00:47:28.630 instead really think carefully about a target of inference

00:47:28.630 --> 00:47:31.220 and then thinking of internal and external validity

00:47:31.220 --> 00:47:34.830 sort of within that and not sort of trying to prioritize

00:47:34.830 --> 00:47:35.993 one over the other.

00:47:37.180 --> 00:47:39.133 And then just an aside, one thing,

00:47:39.981 --> 00:47:42.610 I would love to do more in the coming years is thinking

00:47:42.610 --> 00:47:45.580 about combining experimental and non-experimental evidence.

00:47:45.580 --> 00:47:48.660 I think that is probably where it would be very beneficial

00:47:48.660 --> 00:47:51.780 to go instead of more of that cross designed synthesis

00:47:51.780 --> 00:47:53.083 kind of idea.

00:47:54.810 --> 00:47:57.350 But again, I wanna conclude with this,

00:47:57.350 --> 00:48:00.950 which is gets us back to design and that again, 00:48:00.950 --> 00:48:04.040 sort of what is often the limiting factor here is the data

00:48:04.040 --> 00:48:06.960 and just sort of strong designs.

00:48:06.960 --> 00:48:10.130 So Rubin, 2005 with better data, fewer assumptions

00:48:10.130 --> 00:48:12.980 are needed and then Light, Singer and Willett,

00:48:12.980 --> 00:48:15.680 who are sort of big education methodologists.

00:48:15.680 --> 00:48:19.460 You can't fix by analysis what you've bungled by design.

00:48:19.460 --> 00:48:21.970 So again, just wanna highlight that if we wanna be serious

00:48:21.970 --> 00:48:24.420 about estimating population effects,

00:48:24.420 --> 00:48:26.990 we need to be serious about that in our study designs,

00:48:26.990 --> 00:48:29.610 both in terms of who we recruit,

00:48:29.610 --> 00:48:32.157 but then also what variables we collect on them.

00:48:32.157 --> 00:48:33.070 But if we do that,

00:48:33.070 --> 00:48:36.730 I think that we can have the potential to really help guide

00:48:36.730 --> 00:48:39.380 policy and practice by thinking more carefully

00:48:39.380 --> 00:48:41.843 about the populations that we care about.

00:48:43.020 --> 00:48:44.330 So for more...

00:48:44.330 --> 00:48:46.600 Here's this, there's my email, if you wanna email me

00:48:46.600 --> 00:48:48.500 for the slides.

00:48:48.500 --> 00:48:52.670 And thanks to various funders, and then I'll leave this up

00:48:52.670 --> 00:48:54.560 for a couple minutes,

00:48:54.560 --> 00:48:58.750 which are all big, tiny font, some of the references,

00:48:58.750 --> 00:49:01.060 but then I'll take that down in a minute so that we can see

00:49:01.060 --> 00:49:01.893 each other more.

00:49:01.893 --> 00:49:05.973 So thank you, and I'm very happy to take some questions.

00:49:13.780 --> 00:49:15.500 I don't know if you all have a way to organize

00:49:15.500 --> 00:49:16.400 or people just can

00:49:18.990 --> 00:49:19.823 jump in.

00:49:24.160 --> 00:49:25.200 - So maybe I'll ask the question.

00:49:25.200 --> 00:49:28.003 Thanks Liz, for this very interesting and great talk.

00:49:29.030 --> 00:49:33.500 So I noticed that you've talked about the target population

00:49:33.500 --> 00:49:34.890 in this framework.

00:49:34.890 --> 00:49:39.270 And I think there are situations where the population sample

00:49:39.270 --> 00:49:42.774 is actually a survey from a larger population.

00:49:42.774 --> 00:49:43.607 - Yeah.

00:49:43.607 --> 00:49:46.630 - Cause we do not really afford to absorb everything,

00:49:46.630 --> 00:49:48.750 actual population, which will contain

00:49:48.750 --> 00:49:50.110 like millions of individuals.

00:49:50.110 --> 00:49:54.830 And so in that situation, does the framework still apply

00:49:54.830 --> 00:49:58.370 particularly in terms of the sensitivity analysis?

00:49:58.370 --> 00:50:01.360 And is there any caveat that we should also know in dealing

00:50:01.360 --> 00:50:02.293 with those data?

00:50:03.330 --> 00:50:04.223 - Great question.

00:50:05.150 --> 00:50:07.240 And actually, thank you for asking that because I forgot

00:50:07.240 --> 00:50:09.600 to mention that Ben Ackerman's dissertation,

00:50:09.600 --> 00:50:10.500 also looked at that.

00:50:10.500 --> 00:50:12.920 So I mentioned his measurement error stuff.

00:50:12.920 --> 00:50:16.900 But yes, actually, so Ben's second dissertation paper

00:50:16.900 --> 00:50:20.950 did exactly that, where we sort of laid out the theory

00:50:20.950 --> 00:50:24.100 for when these the target population data

00:50:24.100 --> 00:50:27.033 comes from a complex survey itself.

00:50:28.650 --> 00:50:30.880 Short answer is yes, it all still works.

00:50:30.880 --> 00:50:34.460 Like you have to use the weights, there are some nuances,

00:50:34.460 --> 00:50:36.450 but, and you're right, like essentially,

00:50:36.450 --> 00:50:38.450 especially like in...

00:50:38.450 --> 00:50:41.310 Like for representing the U.S. population, often, the data

00:50:41.310 --> 00:50:44.290 we have is like the National Health Interview Survey

00:50:44.290 --> 00:50:47.040 or the Add Health Survey of Adolescents,

00:50:47.040 --> 00:50:49.110 which are these complex surveys.

00:50:49.110 --> 00:50:52.760 So short answer is, yeah, it still can work.

00:50:52.760 --> 00:50:54.943 Your question about the sensitivity analysis is actually

00:50:54.943 --> 00:50:57.900 a really good one and we have not extended...

00:50:57.900 --> 00:50:59.720 I'd have to think, I don't know, off hand, like,

00:50:59.720 --> 00:51:03.840 I think it would be sort of straightforward to extend

00:51:03.840 --> 00:51:06.560 the sensitivity analysis to that, but we haven't actually

00:51:06.560 --> 00:51:07.393 done it.

00:51:08.340 --> 00:51:09.173 - Thanks Liz.

00:51:10.730 --> 00:51:12.270 The other short question is that I noticed that

00:51:12.270 --> 00:51:16.380 in your slide, you first define, PATE as population ate,

00:51:16.380 --> 00:51:18.650 but then in one slide you have this Tate,

00:51:18.650 --> 00:51:21.150 which I assume is target ate.

00:51:21.150 --> 00:51:24.570 And so, I'm just really curious as to like, is there any,

00:51:24.570 --> 00:51:26.878 like differences or nuances in the choice of this

00:51:26.878 --> 00:51:27.943 terminology?

00:51:28.977 --> 00:51:29.810 - Good question.

00:51:29.810 --> 00:51:30.643 And no, yeah, I'm not...

00:51:30.643 --> 00:51:33.563 I wasn't very precise with that, but in my mind, no.

00:51:34.750 --> 00:51:37.830 Over time I've been trying to use Tate,

00:51:37.830 --> 00:51:39.970 but you can see that kind of just by default,

00:51:39.970 --> 00:51:41.713 I still sometimes use PATE.

00:51:42.830 --> 00:51:45.750 Part of the reason I use Tate is because I think

00:51:45.750 --> 00:51:48.020 the target is just a slightly more general term.

00:51:48.020 --> 00:51:50.210 Like people sometimes I think, think if we meet,

00:51:50.210 --> 00:51:53.330 if we say PATE, the population has to be like

00:51:53.330 --> 00:51:58.030 the U.S. population or some like very sort of big,

00:51:58.030 --> 00:52:00.930 very official population in some sense.

00:52:00.930 --> 00:52:03.570 Whereas, the target average treatment effect,

00:52:03.570 --> 00:52:06.260 Tate terminology, I think reflects that sometimes

00:52:06.260 --> 00:52:10.060 it's just a target group that's well-defined.

00:52:10.060 --> 00:52:10.893 - Gotcha.

00:52:10.893 --> 00:52:12.270 Thanks, that's very helpful.

00:52:12.270 --> 00:52:14.930 And I think we have a question coming from the chat as well.

00:52:14.930 --> 00:52:15.900 - Yeah, I just saw that.

00:52:15.900 --> 00:52:17.450 So I can read that.

00:52:17.450 --> 00:52:19.610 We have theory for inference from a sample to a target

00:52:19.610 --> 00:52:22.700 population needs to find that internal validity approaches,

00:52:22.700 --> 00:52:25.210 what theory is there for connecting the internal validity

00:52:25.210 --> 00:52:26.933 methods to external validity?

00:52:28.620 --> 00:52:32.550 So I think, what you mean is sort of,

00:52:32.550 --> 00:52:36.500 what is the formal theory for projecting the impact

00:52:36.500 --> 00:52:38.110 to the target population?

00:52:38.110 --> 00:52:40.700 That is exactly what some of those people that I referenced

00:52:40.700 --> 00:52:41.533 sort of lay out.

00:52:41.533 --> 00:52:42.366 Like I didn't...

00:52:42.366 --> 00:52:44.590 For this talk, I didn't get into all the theoretical weeds,

00:52:44.590 --> 00:52:46.370 but if you're interested in that stuff,

00:52:46.370 --> 00:52:48.830 probably some of Issa Dahabreh's work would be the most

00:52:48.830 --> 00:52:50.093 relevant to look at.

00:52:51.430 --> 00:52:54.000 Cause he really lays out sort of the formal theory.

00:52:54.000 --> 00:52:58.390 I mean, some of my early papers on this topic did it,

00:52:58.390 --> 00:53:01.220 but his is like a little bit more formal and sort of makes

00:53:01.220 --> 00:53:03.610 connections to the doubly robust literature

00:53:03.610 --> 00:53:04.443 and things like that.

00:53:04.443 --> 00:53:06.040 And so it's really...

00:53:06.040 --> 00:53:08.420 Anyway, that's what this whole literature

00:53:08.420 --> 00:53:11.050 and part of it is sort of building is that theoretical base

00:53:11.050 --> 00:53:12.223 for doing this.

00:53:17.320 --> 00:53:18.503 Any other questions?

00:53:28.070 --> 00:53:28.903 - [Ofer] Liz,

00:53:28.903 --> 00:53:30.226 I'm Ofer Harel.

00:53:30.226 --> 00:53:31.360 - Oh, hi Ofer?

00:53:31.360 --> 00:53:32.670 - [Ofer] Hi.

00:53:32.670 --> 00:53:33.630 (mumbles)

00:53:33.630 --> 00:53:37.453 Just jump on the corridor, so it's make it great.

00:53:39.010 --> 00:53:43.070 So in most of the studies that I would work on,

00:53:43.070 --> 00:53:45.860 they don't do really have a great idea about

00:53:45.860 --> 00:53:50.100 what really the population is and how really to measure

00:53:50.100 --> 00:53:50.933 those.

00:53:50.933 --> 00:53:53.590 So it's great if I have some measure of the population,

00:53:53.590 --> 00:53:57.410 but most of the time it is the studies that I work.

00:53:57.410 --> 00:54:01.630 I have no real measurements on that population.

00:54:01.630 --> 00:54:03.060 What happens then?

00:54:03.060 --> 00:54:03.977 - Yeah, great question.

00:54:03.977 --> 00:54:05.650 And in part, I meant to say this,

00:54:05.650 --> 00:54:07.500 but that's one of the reasons why the analogy...

00:54:07.500 --> 00:54:10.300 Why the design strategies don't always work particularly

00:54:10.300 --> 00:54:12.690 well is like, especially when you're just starting out

00:54:12.690 --> 00:54:13.523 a study, right?

00:54:13.523 --> 00:54:15.973 We don't really know the target population.

00:54:17.070 --> 00:54:21.280 I think certainly to do any of these procedures,

00:54:21.280 --> 00:54:24.840 you need eventually to have a well defined population.

00:54:24.840 --> 00:54:26.950 But I think that's partly why some of the analysis

00:54:26.950 --> 00:54:28.900 approaches are useful is that,

00:54:28.900 --> 00:54:31.090 you might have multiple target populations.

00:54:31.090 --> 00:54:33.010 Like we might have one trial,

00:54:33.010 --> 00:54:35.210 and we might be interested in saying,

00:54:35.210 --> 00:54:38.670 how well does this generalize to the State of New Hampshire

00:54:38.670 --> 00:54:41.370 or the State of Vermont or the State of Connecticut?

00:54:41.370 --> 00:54:45.320 And so, you could imagine one study that's used to inform

00:54:45.320 --> 00:54:47.103 multiple target populations.

00:54:48.050 --> 00:54:49.030 With different assumptions,

00:54:49.030 --> 00:54:50.470 sort of you have to think through the assumptions

00:54:50.470 --> 00:54:51.323 for each one.

00:54:52.390 --> 00:54:53.620 If you don't even, 00:54:53.620 --> 00:54:55.650 I guess I would say if you don't even know

00:54:55.650 --> 00:54:58.560 who your population is, you shouldn't be using these methods

00:54:58.560 --> 00:55:02.040 at all, cause like the whole premise is that there is some

00:55:02.040 --> 00:55:04.900 well-defined target population and you do need data on it

00:55:04.900 --> 00:55:05.930 or at least...

00:55:06.990 --> 00:55:09.340 Yeah, the joint distribution of some covariance

00:55:09.340 --> 00:55:10.380 or something.

00:55:10.380 --> 00:55:13.480 Without that, you're kind of just like,

00:55:13.480 --> 00:55:14.970 I don't know, what a good analogy is,

00:55:14.970 --> 00:55:17.923 but you're kinda just like guessing at everything.

00:55:23.936 --> 00:55:25.650 (mumbles)

00:55:25.650 --> 00:55:27.246 - No, go ahead.

00:55:27.246 --> 00:55:28.864 Go ahead.

00:55:28.864 --> 00:55:30.297 - Oh, Vinod, yeah.

00:55:30.297 --> 00:55:32.380 All my friends are popping up, it's great.

00:55:32.380 --> 00:55:34.370 (laughs)

00:55:34.370 --> 00:55:35.203 - [Vinod] Can I go ahead?

00:55:35.203 --> 00:55:36.923 I feel like I'm talking to someone.

00:55:38.660 --> 00:55:39.980 - Yeah, go ahead Vinod.

00:55:39.980 --> 00:55:42.100 - [Vinod] That was a great talk.

00:55:42.100 --> 00:55:44.320 So I have a little ill formulated question,

00:55:44.320 --> 00:55:47.130 but it's queuing after just the last question

00:55:47.130 --> 00:55:48.956 that was asked is,

00:55:48.956 --> 00:55:53.773 in clinical set populations where,

00:55:54.850 --> 00:55:57.620 in some ways we're using this clinical samples

00:55:57.620 --> 00:56:01.550 to learn about the population because unless they seek help,

00:56:01.550 --> 00:56:05.320 we often don't know what they are in the wild, so to speak.

00:56:05.320 --> 00:56:09.410 And so, each sampling of that clinical population

00:56:09.410 --> 00:56:12.840 is a maybe by sampling of that larger population

00:56:12.840 --> 00:56:14.100 in the wild.

00:56:14.100 --> 00:56:18.450 So I guess my question is, how do you get around this,

00:56:18.450 --> 00:56:21.730 I guess Rumsfeld problem, which is every time you sample

00:56:21.730 --> 00:56:24.140 there's this unknown, unknown, but there's no way to get

00:56:24.140 --> 00:56:27.340 at them because in some ways, your sampling relies on...

00:56:27.340 --> 00:56:29.850 If we could say it relies on help seeking,

00:56:29.850 --> 00:56:33.210 which is by itself as process.

00:56:33.210 --> 00:56:35.160 And if we could just stipulate, there's no way to get

00:56:35.160 --> 00:56:36.270 around that.

00:56:36.270 --> 00:56:38.653 How do you see this going forward?

00:56:39.550 --> 00:56:40.383 - Yeah, good question.

00:56:40.383 --> 00:56:42.650 I think right, particularly relevant in mental health

00:56:42.650 --> 00:56:45.680 research where there's a lot of people who are not seeking

00:56:45.680 --> 00:56:47.106 treatment.

00:56:47.106 --> 00:56:50.090 These methods are not gonna help with that in a sense

00:56:50.090 --> 00:56:53.090 like again, they are gonna be sort of tuned to whatever

00:56:53.090 --> 00:56:54.960 population you have.

00:56:54.960 --> 00:56:56.800 I think though there are...

00:56:56.800 --> 00:56:59.513 If you really wanna be thoughtful about that's

00:57:00.420 --> 00:57:02.870 problem, that's where sort of some of the strategies

00:57:02.870 --> 00:57:05.380 that were used like the Epidemiologic Catchment Area

00:57:05.380 --> 00:57:08.320 Surveys, where they would go door to door and knock on doors

00:57:08.320 --> 00:57:10.660 and do diagnostic interviews.

00:57:10.660 --> 00:57:14.070 Like if we wanna be really serious about trying to reach

00:57:14.070 --> 00:57:16.730 everyone and get an estimate of the really sort of true

00:57:16.730 --> 00:57:20.080 population, then we really have to tackle that

00:57:20.080 --> 00:57:23.253 very creatively and with a lot of resources probably.

00:57:25.027 --> 00:57:26.995 - [Vinod] Thanks.

00:57:26.995 --> 00:57:27.828 - Welcome.

00:57:29.150 --> 00:57:30.430 - Hi Liz?

00:57:30.430 --> 00:57:32.960 Yeah, it's gonna be a true question and great talk

00:57:32.960 --> 00:57:33.793 by the way.

00:57:34.910 --> 00:57:37.576 I'm curious, you mentioned there could be a slight

00:57:37.576 --> 00:57:40.189 difference between the terms transportability

00:57:40.189 --> 00:57:41.070 and generalizability.

00:57:41.070 --> 00:57:42.910 Yeah, I'm curious about that.

00:57:42.910 --> 00:57:45.910 - Yeah, briefly, this is a little bit of a...

00:57:47.563 --> 00:57:48.396 What's the word?

00:57:48.396 --> 00:57:51.120 Simplification, but briefly I think of generalizability

00:57:51.120 --> 00:57:54.670 as one where the sample that, like the trial sample

00:57:54.670 --> 00:57:57.120 is a proper subset of the population.

00:57:57.120 --> 00:58:01.460 So we do a trial in New Hampshire,

00:58:01.460 --> 00:58:04.180 and we're trying to generalize to new England.

00:58:04.180 --> 00:58:07.580 Whereas transportability is one where it is not a proper

00:58:07.580 --> 00:58:10.270 subset, so we do a trial in the United States

00:58:10.270 --> 00:58:12.143 and we wanna transport to Europe.

00:58:13.530 --> 00:58:16.690 Underlying both, the reason I don't worry too much about it,

00:58:16.690 --> 00:58:18.725 the terms is because either way,

00:58:18.725 --> 00:58:20.760 the assumption is essentially the same.

00:58:20.760 --> 00:58:23.130 Like you still have to make this assumption about

00:58:23.130 --> 00:58:25.110 no unobserved moderators.

00:58:25.110 --> 00:58:27.680 It's just that it's probably gonna be a stronger assumption

00:58:27.680 --> 00:58:29.544 and harder to believe,

00:58:29.544 --> 00:58:33.400 when transporting rather than when generalizing.

00:58:33.400 --> 00:58:36.470 Cause you sort of know that you're going from one place

00:58:36.470 --> 00:58:38.053 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.

00:59:37.540 --> 00:59:39.730 But anyway, I think that that is something that like,

00:59:39.730 --> 00:59:41.535 we need a lot more work on.

00:59:41.535 --> 00:59:44.640 One thing, for example, I do have a student working on this.

00:59:44.640 --> 00:59:47.480 Like, we're trying to see if your sample

00:59:47.480 --> 00:59:50.630 is a tiny proportion of the population, like how...

00:59:50.630 --> 00:59:51.670 Cause like there's different.

00:59:51.670 --> 00:59:54.250 That's one where like waiting might not work as well

00:59:54.250 --> 00:59:55.250 actually, who knows.

00:59:56.260 --> 00:59:58.320 Anyways, so like all of these different data scenarios,

00:59:58.320 --> 01:00:00.860 I think need a lot more investigation to have better

01:00:00.860 --> 01:00:03.743 guidance on when the different methods work well.

01:00:09.390 --> 01:00:10.950 Anything else or maybe we're out of time?

01:00:10.950 --> 01:00:13.953 I don't know, how tight you are at one o'clock.

01:00:20.030 --> 01:00:21.980 - I think we're at an hour, so let's...