Cyberseminar Transcript

Date: February 6, 2019

Series: HERC Econometrics with Observational Data

Session: Propensity Scores

Presenter: Todd Wagner, PhD

*This is an unedited transcript of this session. As such, it may contain omissions or errors due to sound quality or misinterpretation. For clarification or verification of any points in the transcript, please refer to the audio version posted at* [http://www.hsrd.research.va.gov/cyberseminars/catalog-archive.cfm](file:///C%3A%5CUsers%5CVHASLCMyersK%5CAppData%5CLocal%5CMicrosoft%5CWindows%5CTemporary%20Internet%20Files%5CContent.IE5%5C73RXHSNF%5Cl)

Todd Wagner, PhD: So it’s a great pleasure to be here today talking about propensity scores. With me is Jean Yoon, so if you submit questions, she is going to be the one who is going to field questions. And Jean, I will ask you, please chime in if you have clarifying questions, or if there are clarifying questions out there. If they are bigger questions, either you can answer them or we can save them till the end.

I have fifty or so slides to get through today. The area of propensity scores is changing rapidly, so when I sort of think about the audience out there and those of you in the audience, if you are highly expert in this area, you are probably going to be a little bit bored by my talk. It is mostly for people who are just starting to get familiar with propensity scores and want more information on it. Like I said, it is growing quickly.

I really have three objectives for you today. The first is to talk about if you are not familiar with what a propensity score is, I want to define it for you and think about it. Then we will do a little bit of discussing about what it means from a randomized trial perspective. We will talk about ways to calculate and implement a propensity score, and then there are some important assumptions needed to make--I keep wanting to say “casual claims” but it is causal claims of observational data. I guess it would be easier to make casual claims.

So in terms of the outline, what I’m hoping to get through today, is I just want to talk about causation and what we mean by causation. We will be using the propensity score as a method to try to think about causation and its limitations and where you are going from that. I will say that there’s been a large growth in prediction models and how propensity scores can aid prediction, without necessarily putting a causal framework on it. That is really not the topic of today. That is really a different set of classes. I have some readings if you are interested in that later. Then the second component is propensity score. We will define it, I will show you how to calculate it, how to use it, and then of course the limitations of the propensity score.

Researchers are often interested in causal relationships. Let me give you some examples of one that we might see in headlines in the news. Does drinking red wine affect your health? Maybe it improves your health or causes you premature mortality. Does a new treatment reduce symptoms? These are all things that we’d like to think of as being causal. Does job burnout affect risk of suicidality? And then you might have some policy questions, maybe the Veterans Crisis Line, does that reduce the likelihood of suicide? These would all be things that we would be interested in understanding from a causal framework. And here correlation may not be sufficient, especially if you are making large investments of money.

Clearly, if you were trying to answer causality, the first approach you that you might think about is what about doing this from a randomized trial. That really is our gold standard for addressing causality, so I just want to provide a framework for how randomized trials do it, and then we will move on to what is happening in propensity score. So the randomized trial provides a really solid methodological approach for understanding causation, and then we will think about how to do that framework within a propensity score.

So here is a diagram. In the traditional randomized trial, you will recruit participants, you are going to do random sorting of those participants into two treatment groups, treatment group A and treatment group B. If the random sorting works well, the only difference between those two groups, A and B, is the treatment itself. So you have a treatment and a comparison and then it is going to lead to different outcomes, so then you can measure the difference in the outcomes as an effect of that treatment.

Now just note that you can, by chance, have poor randomization that can lead to unbalanced groups. Most clinical trials out there that publish these things have a very typical table one that shows the balance across the trials. Many sophisticated trials will implement checks and balances like block randomization in efforts to preserve the randomization and to make sure that they have a better than off odds chance of getting a good randomization.

Just because a randomized trial can speak to causality, you must ask the question for whom and generalizability is often very limited for many of our clinical trials. We end up in a world where we are interested in using observational data.

So in the trial analysis, the expected effect of the treatment, here is some notation for you, is really just the mean difference between group A and group B. The expectation is another way of thinking about sort of the average effect of group A versus group B, and that Y is our outcome. So that could be something like mortality if you are interested in the Veterans Crisis Line. You could also just be interested in something like depression and you have given a new treatment to group A and the usual treatment to group B, and that gives you the mean difference of that randomization. That is what we think of as the mean difference there.

You can take that expectation framework and you can think of it as an equation. A couple weeks ago I gave an overview of looking at equations. If this looks familiar, this is just the equation of a line but we think of it as being a little bit different here because of the randomization and the stochastic nature, so with an error term here. But really, Y is the outcome, A is the intercept, and the difference between treatment A relative to treatment B is that beta coefficient. And here, the i just denotes that in this case the unit of analysis I’ve been talking about is a person, but it could also be groups of people or organizations.

Now you could, if you for example saw slight imperfections in your randomization, you could control for baseline characteristics, and you might do this one for wanting to understand that, but then you could start thinking about expanding this out to look at subgroup effects. Maybe that would give you better handles on mechanisms that would be really important for understanding why. Let’s just say you’re implementing a Veterans Crisis Line and you see overall it had an effect, but maybe it had important subgroup effects that would tell you about why is that crisis line working so well. So you can expand this out not only to control for these things, but then to interact with them.

So there are assumptions that are built into this sort of randomized trial. One of the assumptions is that the right-hand side variables in that equation are measured without noise. They are really considered fixed in repeated samples. If you were to do this trial again and again and again, your right-hand side variables would be the same.

And then there’s an important one here, which is there is no correlation between the right-hand side variables and the error term. What’s so great about a clinical trial is that that is by construction true. So here is a notation for it. Here is the expectation between X and your error term is 0, so this you might see in the literature, it is just another way of restating this assumption. One of the great things about the randomized trial is if these conditions hold, then B is an unbiased causal estimate of the effect of treatment. You get that because you never observe everybody in both treatment groups. You observe people choosing one of the two treatments or getting sorted into one of the two treatments. You have to be careful there. Because it was an external exogeneous choice of which treatment they got, you get to say something about what you think the causal effect is. So this is really the benefit of an RCT, and this is what we don’t get if you are just using observational data, because you often have this assumption does not hold.

Let me just sort of transition now to think about what if. What if the assumptions don’t hold in a randomized trial? Then what? You really lose the unbiased estimate of causality and you start to fail on that regard, and that can be in a clinical trial, too. You could have totally missed your randomization, but this is most often encountered when you are working with observational data.

So think about it this way. There are many people out there who fervently and passionately believe in randomized trials. Not only are they the gold standard for understanding causation, but they think that is really what we should be focused on. But there are times where they are imperfect, impractical, they may not be scientifically justified, unethical. They are also really expensive and they take a long time to do them well. We are working on a trial now that has been 15 or 16 years that we have been working on this trial. It is hard to imagine that the world of cardiothoracic surgery hasn’t changed dramatically in those 15 years, so you start to wonder, are the results--oops, I went to the very end of my slide deck. That was a mistake. Let’s see if I can get to the--alright, sorry about that. Imagine the world of cardiothoracic surgery if it wasn’t for that. I hit the wrong key. Sorry about that.

Alright, so here is a question for you. So when this fails, and we don’t have this exogeneous experimentation, we talk about it being endogenous.

I talked a lot about two weeks ago what this term endogenous is, and I was going to ask a poll just to get a better sense on your perspective on has anyone heard of this term. I have three answers for you. One is: Yeah, I use this term frequently at the dinner table. I will joke that I have two kids, one in high school who is a senior and one in eighth grade, and I will admit that I use this term at the dinner table. So you can laugh at me. And then you can say: Yeah, I’ve heard others use it, but it still confuses me. And then you can say: No, I’ve never heard of this term and I like being honest. So Rob, if you could--so it looks like the poll is open and people are able to submit their answers.

Rob: Alright, the poll is open and people are submitting their answers currently.

Todd Wagner, PhD: There is probably some other answers, too. Yes, I’ve heard of the term but I don’t use it at the dinner table, that is fine too, but I tried to keep this lighthearted.

Rob: Well, things have pretty much leveled off. A little over 80%, so I am going to go ahead and close the poll, Todd, and share out the results.

What we have is that 21% of your audience says that yes, they use the term frequently at the dinner table; 65% say that they have heard others use it but it still confuses them; and 14% say no, and they are honest people. So now we are back on your slide.

Todd Wagner, PhD: Oh, wow. So I don’t know if I should be having dinner with the honest people or the people who frequently use it at the dinner table, like me. That might be a crazy dinner table. But let me go on to another slide where I can talk a little bit more about endogeneity and maybe it will help make a little bit more sense what we are trying to get at here.

So endogenous, it is not attributed to an external factor, so if you look up the root of exogeneity and endogeneity, you are going to see Latin roots internal/external. So let’s think about this example: Does smoking lead to cancer? If you went back 50 or so years in science, there was all these debates about does smoking cause cancer, and it would have been totally unethical to randomize people to smoke or not smoke. So here is a case where you are not going to get a randomized trial of smoking. You are going to have a lot of people looking at observational data and they say, hey we notice that people who smoke tend to get higher rates of cancer, it’s positively correlated. But what is wrong with that? So smoking is also correlated with a whole bunch of other things. Some of these things may be observable, some of them may be partially observable, and some of them may be unobservable to us. Maybe there’s things like your motivation in school, that is typically unobservable. But there could be things like income, education, and what is happening in the household. Maybe if your parents had--if you had talked more about endogeneity at the dinner table, you wouldn’t smoke, and so forth.

So most of the time we can’t control for some of these factors. So by default, when we can’t, and we know that it is partially related to unobserved, you end up with this expectation of smoking and your error term where it is not equal to zero. If you remember a couple slides ago, we talked about what you get in a randomized trial was that this is equal to zero, so here is this problem where it is not, and we end up with endogeneity. Some people might be familiar with the term selection bias. In this case, we’ve got a clear selection bias that is confounding this relationship. So unobserved confounders is another common term for this.

Think about it this way, smoking is endogenous. This is very classic. A couple weeks ago I gave the other example when I gave the smoking, people wanted a different example and I talked a little bit about testosterone or hormones. Many people are familiar with testosterone might be great in terms of building your muscle, and we observe that people with higher levels of testosterone in their body have bigger muscles, but it doesn’t mean that is the same effect when you inject people with testosterone. They’re separate questions and there has to be separate questions of safety.

So let me just go through what I am going to do here as a framework for you. I have a few slides on this and I am hoping to sort of get you to think about this sorting mechanism. In the world where--and I just want to make sure, I borrowed this from Matt and Steve when they gave a Cyberseminar years ago and I liked this framework. So you have this sorting without randomization. This is in the observational data world. You have got this sorting mechanism. There are things that affect people’s sorting into treatment group and comparison group. So think about it, maybe you are interested in smoking and that you are going to say the sorting is into smoking versus nonsmoking, and you are following the outcome. We can observe some of these things, like the patient characteristics, we see that poor people tend to smoke, people whose parents smoked, and you see things like educational attainment is also in there. You might see some geographic areas as well that would fall into that.

If everything is fully observed, if you observed everything in the world and you correctly specify it in your regression model, the results are not biased and you actually end up with consistent results of causality. But that never really happens. You never observe everything. That could be as simple as saying something like motivation for educational attainment, or motivation for more income. There could be a whole bunch of things that you would never fully specify.

So here’s another schematic on sorting without randomization. Here we have the same patient characteristics, provided characteristics, and maybe we have now some unobserved characteristics. We realize there are some things out there that affect outcomes; maybe they are things like teamwork, maybe certain facilities do much better at this. In this case we recognize we have some unobserved that are going to confound us. In some cases, you might be able to say hey, we can observe the facility. So we will just include a fixed effect and we are going to try to take everything out that is observed with that fixed effect. Now that makes assumptions if you use a fixed affect that the unobserved characteristics are fixed, specifically if you are looking over time, you are saying it is fixed over time.

But this is often very typical here, is that the unobserved characteristics are not just associated with the outcomes, they are also affecting the sorting. So here, because you can’t just use fixed effect to say that, so really you have got this much more complicated problem and in the end you end up with this problem where you say causality isn’t identified. You’ll often hear economists talk about what is the identification strategy, identification being trying to understand the causal link. Here, if you can’t observe all of the unobserved and how they affect sorting and outcomes, you can’t identify the causality.

So let me talk a little bit about the propensity score and define it, what it is, how it gets used, and then we will move on from there. What we are going to use to do is we are going to take all of our observed information and think of it as sort of multidimensional. If you’re looking at smoking, think about all the varied factors that could affect smoking--education, parental use, what your friends do--to calculate a single variable. That is going to be the score itself. The score is just the predictive propensity to get sorted, so think of it that way. What we are really trying to do here is predict smoking. That is going to be our key right-hand side variable and so we can think about what are the variables that predict smoking, and that will be our propensity.

So if we work through some math, expected treatment effect here as I’ve shown before, is the expected effect of Y is just the mean difference between groups A and B. The propensity score is just the probability of getting assigned to treatment group A, given your covariates that you have that you’ve observed. We will say that that vertical slash, conditioning on what you’ve observed, what we are really going to come back to time and time again in this model is you’re limited by what you observe. Propensity score is only using observed information.

In another way to say for it, what I like to say is that propensity score really is just another way to correct for observable characteristics. There is no way to take observational data and turn what you think of as a pig’s ear into a silk purse. You just can’t do it. There is no way with a propensity score to adjust for unobserved characteristics. The only way you are going to get there is to make assumptions. I say they are huge assumptions and there’s just scads of literature, huge debates about these assumptions which I will talk about on the next slide and it really is this question of unconfounded or ignorability.

The first question that you are facing to yourself is, are you observing a sufficient amount of information to understand all of the important sorting aspects and such that if you are not observing some things, they really don’t matter. So you have this question of is it really unconfounded? Is it just like a randomized trial? Really here, boy, I guess there is still debate. I would say that the people who say it is not, you can’t take the data and make this assumption is probably winning in this debate. It is a growing field and it is continuing to evolve. I will also say, in the world of Amazon and machine learning, where there is tremendous interest in prediction, there is also a lot of growth in using propensity scores for prediction in which they are not interested in causality. I will say that is a tangent and I will not go into that detail. They use slightly different models in that regard, but it is a separate issue.

Let’s talk a little bit about how do you calculate a propensity score. Maybe you understand that there’s some challenges here with the assumptions of propensity score, I will say that there’s actually some nice things about it, too. I’ll show you how to calculate it and predict it.

So you observe treatment. One group doesn’t receive it and another group does. Let’s just continue our example of smoking. You are going to use a multivariate logistic regression to estimate the probability that you smoke, and think of it as our treatment. Then we are going to take the predicted probability from this logistic regression as the propensity score. Each person is then going to get sort of a predicted probability from zero to one that you are in that treatment or untreatment, or smoking and nonsmoking group.

The variables. There has been a lot of debate about what variables to include in this logistic regression. One of the things that you want to think about is your outcome variable, that is your smoking. What you are hoping to include are certain variables that are correlated with your exposure and your outcome, so I have put these in green as sort of that green is good. You want to include these variables. They will help decrease your variance of your estimated exposure without increasing your bias. You can also think about X2, these would be variables that are just correlated with your outcome but they wouldn’t necessarily be affecting your exposure or the probability of your sorting. You can include those two. Generally, what the literature says--and I gave you the citation below, and I put this in red--is don’t include variables that just affect exposure but wouldn’t affect outcome either. That expands your standard errors if you do do that.

Exclude variables that are related to the exposure, like I just said, but not to the outcome. In practice, it is very hard to come up with examples like this. It is often you find examples that are affected both by outcome and exposure. If you are asked to do this or are doing this, you will want to put some thought into that, and sometimes variables don’t readily come to mind but you can do some testing to figure out what is going on here.

If you do include these variables, you will increase your variance. This is particularly important for small studies, where variance and your efficient estimation is important. That is what the Brookhart says, it gave us a rule of thumb of when you are working with less than 500 people in your sample.

I am going to walk you through an example here. I will take it away from smoking and I will put this back into sort of a health service-y question that we often face. So here is the question; you’ve got a person who is learning how to do surgery. They are a resident. They finish medical school, they are now in residency for surgery, and often they do surgery. The question then becomes: Do bypass patients--this is open-heart surgery--do they fair better or worse when the surgery is done by the resident as opposed to sort of the standard attending surgeon? An attending is, if you are not familiar with that term, is just an existing sort of more senior surgeon.

So we had a dataset that tracked the primary surgeon for heart bypass, and so we had information on when they were done by a resident versus an attending. One of the questions we wanted to understand is: What about the sorting and balance? You could just sort of immediately jump to a multivariate model and say, hey, we are just going to correlate attendings and residents and outcomes and see what happens. But one of the nice things about the propensity score is you think carefully about the sorting itself. We understood that sorting is multidimensional. The propensity score is going to provide a way of sort of reducing the dimensionality so we can think about that and we can adjust to the covariance.

So in this example where we’re looking at whether the surgeon is a resident, we were really worried and talking to a bunch of surgeons, that surgeons and residents are chosen based on certain characteristics that they think is important. They are not going to assign a resident, especially an early resident, to a patient who is very high-risk. So they’ll think about that. The other thing that comes into play is availability of residents. Some facilities have much more residents and other facilities have no residents. There is often this unobserved question about resident skill when you talk to attending surgeons. They know that there’s different residents have skill, and so they are going to not let them do certain things. And there’s obviously differences in local culture.

Let me give you, this is the logistic regression. So I have highlighted some things and I will just walk you through this. This is the odds ratio in the column and then the P-value. We see that age of the patient did not matter here in terms of whether they were getting done by a resident or not. The dependent variable was, did the resident do the surgery is 1, and then if the resident didn’t do the surgery, that’s 0. So we see that the Canadian Functional Class, which is a measure of angina symptoms, so Class 1 is there’s no sort of marked measure of angina or heart pain, if you will. Class 4 is you have heart pain even without doing any exercise. You see that that is markedly assigned to residents, and I apologize, I have the residents flipped here. It is resident is 0 and attending is 1. So that attending is much more likely to do the more severe patients in terms of angina.

And then urgent priority, so some people are wheeled into the OR with urgent, that didn’t seem to matter. You clearly see that there’s some differences across site. We don’t know whether those are due to the availability of the residents, but clearly Site 5 much more likely to use an attending than a resident. And then there is this one that I am going to point out to you, is endovascular harvesting. This, at the time of our data, is minimally invasive harvesting of the vein that you are going to use in a heart bypass and you are pulling this vein typically out of the leg, and so clearly attendings do this much more often. You might have a sense here of imbalance in the sorting. The residents are assigned to patients who the--generally less severe patients, if you will, and that attendings are generally on the main surgery for the more severe.

So one of the nice things--just be careful here--one of the things that I like about this is you think a lot about the sorting. I am going to talk about this term called shared support. One of the questions that we are going to ask ourselves is, can we compare attendings and surgeons? Maybe they are handling such different cases that there’s really no overlap in patients. So if there’s a lot of overlap in patients, that is better. If there is very little, that’s bad. We talk about this as being a common support. I will show this graphically next, but it’s a concept that measures the overlap in sort of a distribution of people. Ideally we’d like to see a fair amount of shared support. And if you see terrible support, sort of common support, you might actually stop the analysis right there and say, we don’t think that there’s sufficient overlap to actually do this with observational data.

So what does this look like? You are going to get two things. This, just to point out, is a kernel density diagram. This is another way of graphing and this was done in Stata. The blue line is sort of the probability of being the resident, the red line is the attending. What I have highlighted here, and I have some--it will be my pen. Oops, I’ve got to select my pen. There we go. So I have done these vertical lines here. I’ve done these vertical lines, man, my mouse skills are just horrible. So you see there are some people who don’t match well in one treatment group versus another treatment group, but more or less they match. Now you can look at the density of the people and it doesn’t match a huge amount.

Two things to be careful of here. What we are looking for is sort of on the X-axis, the number of people who don’t have overlap with the other group. It is not huge in this case. What is more concerning in this case is that we’ve got a huge density in the blue group, in the resident group, that is getting matched to a very small number of people in the attending group. The vice versa is also the case. This might give you some concern about this analysis, that this is not sorting well. What you might want to do--and this is where we’ll talk more about how to actually use the propensity score--you might want to start cutting people, cutting your sample for whom you don’t think that there’s good balance. But hopefully you get a sense on here what we’re looking at is two things on this diagram; one is the sort of shared of the X and the other thing is sort of a density.

I am going to give you another poll here. So here are three propensity scores that we’ve calculated from different studies that I’ve worked on; A, B, and C. I want you to have a question, the next poll is: Is there one of these graphs that is the most concerning to you, given what we just talked about? So let me pull the poll up.

Here’s the answers. You can say: A is the most concerning, B is the most concerning, C is the most concerning, and you might say all of them are concerning, and then you might say none of them are concerning. How about that?

Rob: The poll is up and running and people are providing their answers.

Todd Wagner, PhD: I’m just going to go back in case, I don’t know how the poll looks to them, but in case they want to look at it again.

Rob: They won’t be able to see it until I close the poll.

Todd Wagner, PhD: Okay.

Rob: It’s going a little bit slower this time. It is only around 60% finished, so we will give people a few ore moments to make their choices.

Todd Wagner, PhD: It might be that they can’t remember, so we can move on, that’s okay.

Rob: Okay. I’ll close the poll and share out the results; 76% say A, 4% only say B, 15% say C, 5% say all of them, and 1% say none of them.

Now we are back on those slides.

Todd Wagner, PhD: Awesome. So thank you all for your honesty. In some sense, what I would say, I would be most concerned about A, because if you look at your X-axis here, there’s very few people who share support in this X-axis. The vast majority of the density, here, are not matched to the people over there. You could be concerned about B, too. There’s a definite chunk of the people--oh man, my skills on my mouse are terrible--but you get a sense that those people on the left-hand side of that red line don’t necessarily match up with B. I am a little bit less concerned about C, because yeah, the lines don’t perfectly overlap but they seem to share a lot of the same support, and they seem to share a lot of the same distribution. That one might be a situation where it’s okay. So you end up with a world of, okay so A definitely has a lot of concerns there, B I have some concern, C I am feeling pretty good.

So here’s an example of a situation. What would happen--and this is not a hypothetical, you can do this if you want to if you have randomized trial data. What would happen if you used a propensity score to measure sorting and distribution in a randomized trial? Here I am going to give you two examples of what propensity scores look like in this case. What I’m hoping you see here is that even in a randomized trial there is imperfection. So neither of the densities overlap perfectly, and that is because people differ and they get sorted randomly. What we’re hoping is that there’s not major differences and that is why we do Table 1 in that randomized control trial. So the left graph, you get to see that there’s pretty good overlap. There’s almost nobody who’s outside the bounds of your X-axis, and that is from a randomized trial.

And then on your right graph, that is also a randomized trial. You see that there’s some people in the red graph who are a little bit different, but by and large they are very small in number. The vast majority of the people, the density of the people, tend to be pretty well represented in both. So hopefully you get a sense now, there’s no particular shape that says it’s a good shape or a bad shape. We’re looking for the X-axis and then we’re looking for the probability match between the two of them.

So there are some evidence that propensity scores provide advantages when there is considerable shared support. One of the reasons I like this in the teams that I work with, is it gets people to think about the sorting. For one thing, if you went back three slides and said that graph A, where there’s very little, you might say hey, let’s stop here. We’re just never going to get good balance to go further. And I actually saw a paper like that. So to give kudos to Matt Maciejewski at Duke and the Durham VA, was working on this one with I think it was gastric bypass, and they didn’t support, good matched support, and they called it a day right there. They said, we can’t make further evidence here, we can just do correlational stuff but clearly these people differ. And then there is growing evidence that it can create problems if you’ve got poor support.

So let’s talk about how do you use the propensity score. So at this point in your mind you’ve figured out that you’re going to run a logistic regression model, you’re going to save the predicted probabilities, so now you have this predicted probability. What are you going to do with it? You’ve graphed it out and you’ve looked at your shared support. Well maybe you can do something like, hey, let’s compare individuals based on similar propensity scores. There’s different ways to do that, but you can come up with matches, right? So if you, sort of a Euclidian space, you might say let’s find the nearest neighbor to this person. What you are thinking in your head is maybe if they match on observables, they’ll match on unobservables, too. Maybe you can do some sub-group analyses and stratify your analysis to say, hey let’s look at people who match at this point on the shared support, this point on the shared support, and so forth. Maybe that gives you information on how the propensity score is working on mechanisms.

You can include it as a covariate, whether it is the propensity score itself or is a quintile. Most often it is some sort of transformation like a quintile in the regression model. There’s been some work on weighting the regression model, so here’s you’ve got your outcome model, let’s say it’s smoking again or you’re looking at mortality and residents. Now you can maybe put more weight on the similar cases, so use the propensity score as a weight.

Then there’s another method that came up sort of well-known about a decade ago called doubly robust, which uses both three and four.

So let’s talk about each of these a little bit. Using it as a covariate doesn’t seem to offer many advantages, and most people just don’t do that and shy away from it. This the idea was you calculated your propensity score, now in addition to your regression model looking at let’s just say it’s smoking, you throw in your propensity score. It doesn’t add a whole lot there. One of the things that propensity score does do if you break it into quintiles, it adds a little bit of flexibility in your functional forms. Things like interactions and nonlinearities and so forth. And then if you’re working in really small samples, there’s perhaps some evidence to say that the propensity score offers some advantages to efficiency when you are working with really small samples in this regard.

A lot of people use it for matched analysis. What they’re really thinking, as I talked about, is that you’re thinking hey, we’re trying to match people as close to possible on an experimental person versus a control person, hoping that this somehow matches them on the unobserved too. In this case you would exclude cases and controls that just don’t match and in some sense don’t share any support. You end up reducing your sample a little bit, which might reduce your generalizability, your power, and there’s a whole host of different matching models out there. I have some literature on that if you’re interested in sort of matching analysis, as well as a recent paper by Gary King that sort of highlights some problems with matching.

One of the matching methods, and I’ve mentioned it, is just the nearest neighbor. You are going to take the propensity score and you’re going to come up with some sort of distance measure and you are going to choose the nearest neighbor. What I like about this is that everybody has a neighbor around where you live and you can think about who is my nearest neighbor. This isn’t the neighbors around you, this is sort of the neighbors who share your similar characteristics. It is trying to find the person who shares the most similar characteristics as you and then use that as your comparison. Obviously if you’re in the experimental arm, that would be the most nearest neighbor to you that’s in the control arm.

There’s another technique called caliper. Again there’s a bunch of different methods here that people have gone into many details on, where you’re trying to not choose one but maybe do random draws here.

Then there’s evidence that the choice of what you match in your estimator here is important. I will say for those of you who start running these models, is that most papers nowadays, if you submit a paper to a journal with propensity score on it, they are going to want to see some supplemental evidence in a technical appendix. You are going to start walking people through that what you chose was the right thing, that it made a difference, and that there wasn’t a better choice.

So recent areas of research on this, it’s coming really fast and furious from different areas. In economics, I heard Susan Athey speak yesterday. She just gave this amazing talk. She is a professor of Econometrics here at Stanford, and she was talking about machine learning and using machine learning for understanding causal identification. She has done some great papers looking at propensity scores and she was one of the people talking about the propensity score estimates for prediction models. She’s got some great work there.

Gary King, this is still a working paper; he is a political scientist at Harvard. You can tell by the title there, Why Propensity Scores Should Not Be Used for Matching, immediately what his take is; I would encourage you to download that. Essentially he is saying that we can get really biased estimates. Almost everybody who finds a paper that says, hey propensity score is the greatest thing since sliced bread, you’ll find a similar paper that says we’re running the risk of destroying humanity with it.

Where we’re seeing sort of in biostats and in medicine is these high dimensionality propensity scores. Think about it this way; in VA, if you’re working with VA data, you might have a lot of information about people that you could match on. Sooner or later we’re going to have genomic data on people. Maybe we wouldn’t do our matching just on how they use the healthcare system, we could actually start matching people on genomic data. So that’s where the sort of biostatistics is running to say, what if we get into these really high-dimensional propensity scores, do we get any benefit from doing that. The field is just going nuts. I think that is a good thing.

There are some limitations. Let me walk you through the limitations.

The first is the question that we get back to, this assumption of unconfoundedness and do the unobservables matter? I have to take a little bit of a break here because it is my bias. I think it’s almost everybody’s agreement is that unobservables matter and you can’t just brush them under the carpet. So propensity scores are a great method for using observable data, perhaps more observable data in different ways, but it does not--unless you’re willing to make an assumption--it does not condition on unobserved. There’s no way for it to affect unobservables. It’s really improbable that we observe this sorting mechanism in every detail, so that it’s fully observed, to make assumptions there. So in that regard, we are still back on this assumption, we probably still have endogeneity, we probably still have biased estimates that say nothing about causality. If you believe that, then you have to sort of figure out how to get back to thinking about identification and then you are going to start getting back into maybe there are other methods.

Just a plug for some of the future talks that we are going to be giving; one on instrumental variables, we talked on fixed effects, we’ll be talking about regression discontinuity. I think my RCT there should be regression discontinuity. Different ways of thinking about sort of policy evaluations and rolling them out so that we can understand more about the unobservables and making causal claims.

Here is a really interesting set of questions. Maybe propensity scores, in the best of scenarios, just makes thing a little bit better. Maybe in the worst scenario they actually make things worse. There are two papers here. The Brooks and Ohsfeldt paper came out a couple years ago, 2013, Health Services Research, where they actually simulated the data. One of the benefits you get when you simulate the data is you know exactly the underlying data generating process and you can create things and then say, hey we are going to create this dataset where we have observed different things. We know the mechanism of relationship between these variables, and then we are going to assume some of those variables go unmeasured. We are then going to constrain our simulation on some subset of data and what happens to those unobservables. What they say, you can look at their title, they said: squeezing the balloon. So it’s a great analogy. What they said is when you start to constrain your data and do all this matching, you can actually cause greater imbalance in your unobservables.

Again, it’s a kudos to using simulated data. In most real-world applications, because it’s not simulated data, you don’t know the effect that you’re having on the unobservables. What I tend to see in most papers that I review and that people publish is they’ll walk through the different steps. They’ll show sort of the unadjusted data. They’ll show the adjusted data, maybe from a multivariate regression model, and then they’ll show the propensity score matching model, for example. You might be able to walk across and say, hey the results are consistent across the multivariate and propensity score. We just had a paper, or we’re working on a paper, where we were looking at regression and propensity score. In the regression we were getting this marginally significant effect and in the propensity score we didn’t. And so people were like, what does that mean? My interpretation of that is when you have better control of the data, that effect went away. So that didn’t bother me. That is in a case where I’m just sort of like, yeah so that correlation doesn’t really exist. We don’t have to worry about it. But I’m still not making causal claims there.

Then there’s this great paper by Gary King talking specifically about matching and the potential biases that matching can cause. You can go right to that link. You can also just, if that link doesn’t work, you can then just type in Gary King propensity scores and you’re going to get to his Harvard website. You can just download his paper.

So just to summarize, because we’ve covered a lot of territory. Propensity scores offer another way to adjust for confounding by observables, but they are not a method for measuring unobservables. You can make assumptions that if you think you’re doing a better job about measuring the observables that you might be doing a better job on the unobservables, but I just gave you two papers that suggest that is not necessarily true. One of the nice things about propensity score is that it reduces this very sort of multidimensional problem of understanding the sorting into a simpler way to understand sorting. I find that to be actually helpful. When I went back and we do that residency/attending surgeon, you can see exactly how the sorting worked. You can say, wow, residents are really working on simpler surgical patients.

Then just to highlight, especially if there’s methodologists out there, there’s many ways to implement propensity scores. A growing interest in matching estimators, as well as a growing concern, and that’s the Gary King paper.

One of the things that I do like about propensity scores is it allows for people to observably check between control and treatment balance. Just to be careful though, we’re still making assumptions there, so you just can’t assume everything, but just without balance, and I gave a shout out to Matt and his team. One of their studies they just stopped and said we can’t actually do much more than this because we have so little shared support. But we just have to recognize that the average treatment effects may be very sensitive to the choice of the estimators that we use. If you are not familiar with Guido Imbens’ work and Jeff Wooldridge’s work, here’s a link. There’s a series of lectures that they go through on the NBER, the National Bureau of Economic Research, that are actually just great estimators--great set of papers, sorry. Terrific papers.

Challenges. The first challenge is I think that people misunderstand what a propensity score is, and that is still very common. I see people thinking, oh they are so much better than multivariate, we can just use propensity score and then we can say something about the causality. No, that’s not exactly what you can do here. I’m hoping that I’ve demystified exactly what propensity scores are, and sort of the challenges therein. I will say that most people, let’s just say I’m working on a large team and people are asking to do different components. We will talk about different methods for handling it. People will say, oh just run a propensity score. There’s just not one propensity score. There’s many ways to run a propensity score, and the different methods matter. I do worry that people aren’t placing enough attention onto the estimator itself and the robustness. You start getting into questions of should we be dropping people, are we matching it; if we are matching, how are we matching. I would encourage people to think carefully about these things, and I think there’s a desire just to sort of rush through them. That’s again the robustness checks.

And while we can create perhaps better balance on observables with propensity score than we can on just multivariate regression, there’s--I’m still worried about the unobservables and the selection bias that goes into it. So just because we can have control for differences between residents and attendings, I don’t think they’re the same. I wouldn’t, if I was suggesting a very sick relative to get surgery, I wouldn’t suggest that it’s indifferent or he or she be indifferent between a resident and attending. I am not going to fall into that trap there.

The future reading is just overwhelming. First and foremost, you have to be familiar with Rosenbaum’s and Rubin’s paper. This is sort of this 1983 paper, got it all going, and it is just sort of the classic in the field. But then you get that series from Imbens and Wooldridge; if you’re not familiar with Jeff Wooldridge, he’s at Michigan State and just is a phenomenal econometrician, does a great job on especially panel data. You see a bunch of stuff there from Guido, and then you get into Gary King’s work is there, the Ohsfeldt paper is there. If you’re in VA, you get to see Melissa Garrido, she’s at PEPReC in Boston. She’s doing some studies on propensity scores and some other work. I had a citation earlier for Susan Athey’s work. I didn’t get it into this reading list, but she’s been doing some great work too.

So, questions? I’ll open up the floor. Just a shout out to Jean, who is going to present our next class on February 13th on Difference-in-Differences. We try to promote our classes on Twitter, so feel free to join HERC or myself on Twitter, and any questions there.

Jean, there’s probably been more than one question. I apologize.

Jean Yoon, PhD: Yeah, there is. A couple of questions are in the queue. If we don’t have time to get through all of them I encourage you to contact Todd or HERC directly. The first question asks: After using propensity scores to exclude the cases in the final regression model that is predicting the outcome, do we still put the treatment and other covariates in the model?

Todd Wagner, PhD: That’s a great question. For most of these things, what you’re going to find is there’s not a cookbook. There’s no simple way of saying, hey here is how you do it. If I give you my rules of thumb, it’s different from what you might see elsewhere. So generally speaking, what you’re going to do is you’re going to have, let’s just say you ran a matching estimator, you looked at your shared support, you realized that some people didn’t have shared support. You kicked some people out, so you shrank your sample size a little bit. You’re then going to do matching between a person, a case and a control. You don’t have to put your other covariates into that model. You can just do a matched comparison at that point, because they’ve been matched on propensity. You can add the other covariates, they just don’t add a lot because of the way that you’ve already structured your data. In many cases, you’ve put in all that information into the propensity score already.

Jean Yoon, PhD: Okay, great. The next question asks: What evidence do you use to demonstrate the quality of the propensity score match?

Todd Wagner, PhD: Quality is the eye of the beholder. So there are tables that you will see that sort of look at the sort of improvement in fit for propensity scores, standardized differences and so forth. There are some rules of thumb. It all comes down to what is your intuition and sort of what are you willing to accept. What I particularly like doing is graphing the propensity score, just seeing how much shared support I have. That is why I sort of spent so much time on that, to think about do I really think I should be comparing these people. But there’s no necessarily great answer to say you’ve got a high quality propensity score, you could try it different ways, I guess that’s sort of the art of doing this, or sort of the art and the science of doing this, versus just there’s an easy answer for you.

Jean Yoon, PhD: Okay. The next question asks: What percent of cases should be matched to controls to get an unbiased result? What should I do if I get only a small proportion like less than 50% match?

Todd Wagner, PhD: Great question. So this person sounds like they are running a matching estimator and they have low shared support. So there is no threshold by which you would say X number is bad or good. I think you have to answer that for what makes you comfortable. It sounds like you’ve got a situation where if you were to graph out your probability distribution functions, you’re probably not seeing a lot of shared support. The question I would pose to you is, does that make you feel uncomfortable, and it sounds like it does. I think your concern is justified. The question then is, do you stop then and say we can’t find good support, or do you then just focus on the people for whom they share and you have good matches. I would think in practice, if your sample size is large enough, you might start with that as a first step and say is that consistent with what I get when I do an overall analysis. But if you get two totally different analyses or answers, you might then say we’re getting very different answers, we probably can’t go further here.

Jean Yoon, PhD: Okay, great.

Todd Wagner, PhD: And just to follow up there, I should say these are my gut reactions to the answers, Jean. Feel free to chime in. I think whenever we do analyses like these, what we’re trying to be is as transparent as possible to the audience. If you have grave concerns about your analysis and they really change depending on what covariates you put in your model or which method you use, I think that’s incumbent upon you to let your audience know that. Because that’s part of the research, too.

Jean Yoon, PhD: Okay. Here’s a question, so it asks: When we critically appraise the quality of a given study that uses propensity scores, what should we be looking for?

Todd Wagner, PhD: That’s another great question. So one thing is it’s almost impossible for them, with journal limits, and even with supplemental tables, to put everything into it that might be useful for you. In many regards, I think you can also just contact them if you have questions. But hopefully they have provided some technical appendix, because it’s typically not in the article itself, that’s going to show or talk more about the shared support, the shared differences, how much the fit improved when you did propensity versus didn’t do propensity. Perhaps they tried different propensity score methods and maybe they were all consistent and they chose the easiest, but I think you’re going to find some of that in a technical appendix. The understanding of the quality though, is really tough. What you’re not going to get at is the concerns that Ohsfeldt and Brooks had or what Gary King had, which is maybe this propensity score is making the situation worse. There’s no unobservables. It’s not simulated data, so there’s no way to know with a gold standard that you made it better versus made it worse.

Jean Yoon, PhD: Okay, great. The next question asks: I’ve noticed in drug A versus drug B comparisons, you often have a treatment versus control issue where the average control is more acutely ill than the average treatment. So when you match you end up with an overall lower acuity of patients. Is this a common problem with propensity score?

Todd Wagner, PhD: So it is, and it is just like that resident versus attending. For a whole bunch of reasons, people who seek treatment and are willing to go to take additional medications are typically more severe than patients with more mild cases. I don’t what illnesses people might have, but even think of a situation where it’s just like you have mild hypertension. The question would be, at what level of hypertension do you want to take medication. Not everybody wants to take medication, medication costs things, there’s copays, so you’re going to find a natural sorting. Then you’re going to have to ask the question of can you make a good comparison across this gradient of acuity. This is why the FDA will not accept observational data for drug approval. They still require randomized control trials. It’s because they’re too worried about the unobserved, there. So you might end up with data that is trying to do that. What I would say is maybe there was some natural experiment or shock that induced people to get treatments or get more medication who weren’t on medication or vice versa. So when we think about maybe there are ways to look at identifying causal effects where you’re not so worried about this gradient acuity.

Jean Yoon, PhD: Okay. We have about two minutes left. Do you want to take another question?

Todd Wagner, PhD: Sure. We’ll take one more and I’ll apologize if--sorry.

Jean Yoon, PhD: There were a couple of people asking if you could explain shared support.

Todd Wagner, PhD: Alright, so shared support. Support is a statistical term, but if I come back to here, what we’re going to find is that on a single dimensional level the shared support is the X-axis for which patients in treatment group A, this resident versus attendings, where there are people in the other group. I’m not saying that particularly well. Where there are controls and experiments on the same covariates, is one way of thinking about it. Now what I was worried about in this, we have quite a bit of shared support, I would say. It’s like 95% shared support. I drew the red lines where outside of those bounds there’s not shared support. You can immediately imagine kicking those folks out. The other thing that might give you concern here is that the vast number of people in the blue arm are getting matched to a very small number of people in the red arm, and vice versa is true. This might be a situation where you’re saying, what happens if we come along here and really wanted to focus in on this group where we have a much higher shared support and it looks like we’re doing a better job matching. Now of course, you’re sort of throwing all these folks out and you’re throwing all these folks out, so your efficiency and your standard errors, your efficiency is going to crash and your standard errors are going to go way up. You’re going to have much smaller people, but you might feel like these people might be in the middle, might be much more sort of comparable, if you will. So the shared support is just sort of shared on the X-axis, here. I apologize for running out of time.

Jean Yoon, PhD: There was a little bit of confusion about what the X-axis was. That was the propensity score going from 0 to 1?

Todd Wagner, PhD: That’s correct. So that’s yes, you’re graphing the propensity score for each group. I apologize if that was not understood. So this is, if you’re down here in this blue bar, you have a low propensity to be an attending, because you’re very close to--you’re in the residents group, and then vice versa. It’s showing your probability of being an attending.

Jean Yoon, PhD: Okay, we’re past noon. Did you want to answer any questions?

Todd Wagner, PhD: I think I’ll probably have to let people go at this point because we promised to let them go and then they can contact us if they have further questions about how to do this.

Jean Yoon, PhD: Okay. So sorry if we didn’t have time to get to your question, but feel free to contact HERC directly. And Rob, do you want to say any final words?

Rob: I do. Maybe if people don’t ready know it, what is the correct email address to contact HERC directly?

Todd Wagner, PhD: HERC@VA.gov, so it’s the first thing on that slide is the best way to contact us.

Rob: There you go.

[ END OF AUDIO ]