# Using Propensity Scores in Causal Inference

### with Donna Coffman and Max Crowley

### Donna Coffman

### **April 13, 2012**

Host [Aaron Wagner](http://methodology.psu.edu/people/awagner) interviews Methodology Center Research Associate [Donna Coffman](http://methodology.psu.edu/people/dcoffman) and graduate student [Max Crowley](http://methodology.psu.edu/people/dcrowley). They discuss using propensity scores for causal inference. This is also the topic Donna will present at the upcoming [2012 Summer Institute on Innovative Methods](http://methodology.psu.edu/summerinstitute/). Propensity scores allow researchers to determine cause and effect in experiments that were not randomized.

**Podcast Timeline:**

00:00 - introductions

01:17 - overview of propensity scores

05:53 - a hypothetical example / including confounders in the model

11:46 - ways to use propensity scores

17:29 - resources for using propensity scores

Speaker 1: Methodology Minutes is brought to you by the Methodology Center at Penn State, your source for cutting-edge research methodology in the social, behavioral, and health sciences.

Aaron: Hello and welcome to Methodology Minutes. Today we'll talk to Donna Coffman and Max Crowley about estimating causal effects. Donna and Max will be facilitating this year's Summer Institute on innovative methods where they will be talking about the same topic we are discussing today, propensity scores.

Donna Coffman is a research associate at the Methodology Center and a Research Assistant Professor of Health and Human Development at Penn State University. She received her Ph.D. in quantitative psychology from the University of North Carolina Chapel Hill. Her research interests include extending propensity score methods for causal inference.

Max Crowley is a prevention and methodology Pre-Doctoral Fellow at Penn State University. His research interests include applying propensity scores through the analysis of prevention programs.

Donna and Max, thank you both very much for being here and welcome to Methodology Minutes.

Max: We're happy to be here.

Donna: Thank you.

Aaron: When we talk about causal effect, what would normally be done? In prevention research, how would someone generally measure causal effect?

Max: In general, randomized control trials are often the study design that most people tend to think about when they seek to estimate the causal effect of some sort of variable. Sometimes randomization is not possible.

Aaron: How does that relate to this year's Summer Institute on propensity score methods for causal inference?

Max: In certain cases it's not possible to randomize individuals to certain levels of variables. For instance, in the case of researchers who are interested in how parenting behavior leads to later substance abuse, it's pretty much impossible to randomize parents to deliver certain types of parenting to their children essentially. The result here is that it's very difficult to look at the actual causal effect of parenting on later outcomes.

Another example would be that sometimes certain treatments that are being delivered in real-world settings cannot be randomized for ethical or political reasons. That often results in situations where there's no control group and no randomization has actually occurred, again, masking potential causal effects.

Aaron: What sort of thing couldn't be randomized like that for, say, a political reason? Give me an example of that please.

Max: There are situations especially around health and behavioral medicine where individuals cannot be randomized to a pure control group. They often are required to receive some level of standard of care in order to meet the ethical guidelines of that organization.

Donna: For example, you're looking at treatments for suicide prevention. You cannot not give the treatment to someone who is contemplating suicide. Everyone has to receive some type of treatment.

Aaron: What can researchers do when randomization is impossible or impractical?

Donna: Essentially the issue is that there are variables that impact both selection into the treatment and the outcome variable. These are called confounders. Traditionally confounders have been controlled by adding them as covariates in a regression equation but there are other methods of controlling for confounding and selection bias besides including the confounders as covariates in a regression.

One way is to use propensity scores. Propensity scores are based on the potential outcomes framework for causal inference. Under the potential outcomes framework there is a potential outcome for each individual under each level of the treatment. For example, each individual has a potential outcome that would occur if they received the treatment and they each have a potential outcome that would occur if they received the control group.

The causal effect is the difference between these two potential outcomes. Of course, in reality only one of these two potential outcomes is ever observed. For an example, if someone's in the treatment group then we can observe the outcome that would occur if they received the treatment but we wouldn't be able to see what would happen had they not received the treatment.

An often-used example of this is, if you have a headache and you take an aspirin and in an hour your headache goes away you don't know if it would have gone away anyway because time has passed or if it's because you took the aspirin. You don't know what would 5have happened had you not taken the aspirin.

That's kind of the framework for causal inference that propensity scores are based on. Basically propensity scores control for confounding by trying to model the selection bias. For example, propensity scores are defined as the probability of receiving a particular level of the treatment, given all of the measured covariates.

You can contrast that with a regression adjustment where, in the regression model, the measured covariates are used to predict the outcome. In this case the measured covariates are used to predict the selection into the treatment.

Aaron: Can you give us an example of something that can't be randomized but where propensity scores can help you determine causal effect?

Max: Many state governments now are implementing home visitation programs within their state using federal monies but many of those states are hesitant to use any of that money to set up control groups so that they can evaluate those programs down the line because that would require essentially not giving individuals who need those home visiting programs those programs. That often can be an ethical issue as well as a political issue.

Aaron: There was no control group for not receiving home visitation, so how do we tease out what impact these home visits had since it's everybody who received the program?

Donna: Yeah, you would have to find people in the general population who are like those people who receive home visitation.

Max: We can do that by using different propensity score methods to create equivalent groups, just like we would in the context of a randomized control trial. In a randomized control trial we randomize and that gives us balance in equivalent groups. Using propensity scores we can achieve a similar effect.

Aaron: If I'm understanding you correctly using the confounders helps you build the propensity score models, is that right?

Donna: Yes, that's right. Basically you predict the probability of receiving the treatment, in this case the home visitations given all the potential confounders or all of the measured covariates that could be potential confounders. The way that propensity scores are most often estimated is to use a logistic regression and in this case the treatment versus control is the outcome variable.

All of the measured covariates are predictors in this logistic regression. Then the propensity scores are the predicted probabilities from the logistic regression model. These predicted probabilities are used as estimates of the propensity score. It's the propensity to receive the treatment.

Aaron: How do you know which covariates to measure, which confounders to include in the model?

Max: What can guide that decision-making process is similar to what has guided the inclusion of different confounders as fixed effects in models previously, but as we know from recent studies just including things like demographic factors as confounders is not adequate to account for all of the selection effects, all of the bias, that could be coming from not having a randomized group or not having equivalent groups.

Beyond just demographic variables we can also think about variables that specifically are related to, or correlated with, both the outcome of interests as well as the treatment, so whether or not they received the treatment as well as related to the outcome in the future.

Donna: Most certainly you want to include variables that are related to both the treatment and the outcome. One type of variable that you don't want to include in a propensity model is a post-treatment variable, or a post-treatment confounder, so in other words a variable that may have itself been influenced by the treatment. In other words it's an outcome of treatment. You don't want to include something like that.

For an example, if you're looking at the effect of an alcohol treatment program on later relapse and in between the time that the treatment is delivered and relapse is measured, some of the people are institutionalized, then this would be an example of a post-treatment confounder. That is the type of variable that you would not want to include in a propensity score model.

Aaron: In the hypothetical example of a home visitation program what are some of the confounders that would exist in that situation? What do you measure so that you know who in the general population you can compare to the home visit group?

Max: For many home visitation programs one of the criterion for being in the programs is the mother's risk status. Often if she's a first-time mother or young or living in poverty, unemployed, those are factors related to whether or not the mother could actually receive the home visiting program.

In the simplest case, then, that would be one of the factors you would want to include in your estimation of your propensity scores because you wouldn't want to compare individuals, mothers, who would never have received the program because they didn't have those risk factors.

You would want to make sure that you were excluding individuals from your comparison group that weren't at risk and would have never received the program, otherwise your control group could look dramatically different than your treatment group, the group receiving the home visiting program.

Donna: This brings us to a very important assumption of using these propensity score approaches. That is, the propensity scores are based on the assumption that all of the potential confounders are included in the propensity model, or another way of saying that is that there are no unmeasured confounders.

When designing the study you want to collect as much information as you can on potential confounders or variables that you think may influence both selection into the treatment and the outcome.

Max: Because if you don't then you're violating a key assumption of the larger potential outcomes framework that you have no missing confounders essentially.

Aaron: The score is a measure of how likely you are to undergo a certain condition, to receive a certain treatment, to receive a home visit in the example that we gave earlier. What does the researcher do once he or she obtains that score?

Donna: There are various methods of using the propensity scores once they've been estimated. The three general categories of approaches are matching, sub-classification, and weighting. In matching you essentially match someone who's in a control group with someone who is in a treatment group who has the same, or very nearly the same, propensity score.

In weighting essentially what you are doing is you are weighting each individual by the inverse of their propensity to be in a particular level of the treatment. Weighting approach is similar to survey sampling weights.

In survey sampling the goal of using the weights is to weight the data that you have, weight your sample, to mimic, or be representative, of the population. Here the idea is very similar in that the weights are essentially re-weighting your observed data to mimic what would have been obtained from a randomized trial.

Sub-classification is essentially putting individuals with similar propensity scores into a subclass or strata, and then within each of the strata of similar propensity scores estimating the causal effect of the treatment on the outcome.

Basically in any of these three approaches to applying the propensity scores the researcher is controlling for the confounding by changing the data. For example, if you use matching then once you have obtained matches you then estimate the causal effect of the treatment on the outcome within that matched data set.

Likewise if you're using weighting you estimate the causal effect of the treatment on the outcome in the weighted data set. In contrast to regression-based approaches where you control for confounding by adding covariates to the regression model, in the case of propensity scores you're using the propensity scores in some way to control for confounding by changing your data, by changing your data to mimic what would have been obtained in a randomized control trial.

Aaron: In matching you would find someone who had the risk factors, the confounders that we mentioned earlier: their level of poverty, being a first-time mother, being very young. You find someone in the treatment group and you would match them with someone in the general population. Is that right?

Donna: Yes. The theoretical property that allows propensity scores to work is known as the balancing property. What that says is that an individual in the treatment group and an individual in the control group who have similar propensity scores are also equivalent on all of the confounders that were included in the propensity score model.

This assumption can be verified by assessing balance. Basically this is done by computing the standardized mean differences between the treatment and control group on all of the measured confounders that went into the propensity score model. Balance is assessed both before and after the matching procedure.

Of course, most often before the matching procedure there's going to be medium, large, fairly large, standardized mean differences between the two treatment levels on these confounders. You should see that all the standardized mean differences after the matching procedure are less than the absolute value of 0.2, which Cohen suggested that a 0.2 standardized mean difference is essentially a small effect size. After the matching procedure all of the standardized mean differences are less than small.

Max: Randomized control trials achieve this balance through the randomization process, which is why they are often considered the gold standard but that, of course, leads us back to situations where randomization is not possible.

Aaron: You mentioned three methods for applying propensity scores. How does a researcher know which method is appropriate to use?

Max: This is actually an area where there's a lot of discussion around, what's the best approach for applying propensity scores to different situations? There's not a clear consensus around it, although in specific cases sometimes different approaches are more useful.

Donna: Yes, for example with time-varying treatments it's very easy to extend weighting to that context, where it's not as easy to extend matching to evaluate the causal effect of time-varying treatments. In other cases matching might be a better approach to take particularly if the propensity scores are close to either zero or one because when this happens the weights will get very, very large or very small.

When the weights get very, very large you could, for an example, have one person representing 250 or 300 people. In this case the weights are just too extreme and matching might be a better approach. It really just depends on the situation as to which approach is best.

Aaron: You've set me up to ask you a real softball question. Is there any sort of convenient training where researchers could come and learn about these propensity score methods and which to apply and how in greater detail?

Donna: Yes. This summer I will be teaching the Summer Institute here at Penn State. We will talk in a lot more detail about estimating propensity scores, applying them, selecting confounders to include in the propensity score models, how to assess balance and we'll move into some of the more advanced applications of propensity scores such as evaluating the causal effect of time-varying treatments. We'll move into more complex applications of propensity scores including time-varying treatments.

Aaron: Great, just to complete that shameless plug, you can register for the Summer Institute by going to our website methodology.psu.edu. In the meantime what resources currently exist for researchers who are interested in propensity scores?

Max: That's a great question. For those who are interested in more of an introduction around potential outcomes framework and causal inference in general we definitely recommend Rubin's 1974 Seminole paper published in the Journal of Educational Psychology.

For those who are seeking a little more in-depth reading around estimation and matching we recommend Rosenbaum and Rubin's 1985 paper published in the American Statistician. Further, for those who are interested in sub-classification, which Donna just mentioned, we recommend Rosenbaum and Rubin's 1984 paper published in the journal of the American Statistical Association.

Donna: Finally, there is a tech report, Kaufman, Moore, and Lanza on our website where we used each of the approaches for applying propensity score methods in one data set and looked at the comparison of those, so that's available for download too.

Aaron: Full references for all four of those papers are available on the webpage where you downloaded this podcast. Donna and Max, we're looking forward to the Summer Institute and we'd like to thank you for joining us today.

Max: Our pleasure.

Aaron: Thank you.

Speaker 1: You have been listening to Methodology Minutes brought to you by the Methodology Center at Penn State, your source for cutting-edge research methodology in the social, behavioral, and health sciences. This podcast is available on iTunes and at methodology.psu.edu.

**Selected Articles Related to this Podcast**

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology, 66(5), 688-701. [View abstract](http://psycnet.apa.org/index.cfm?fa=search.displayRecord&uid=1975-06502-001)

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. Journal of the American Statistical Association, 79(387), 516-524. [View abstract](http://www.jstor.org/stable/2288398?origin=crossref&&)

Rosenbaum, P. R., & Rubin, D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. The American Statistician, 39(1), 33-38. [View abstract](http://www.jstor.org/stable/2683903)