Assessments of Causal Effects—Theoretically Sound, Practically Unattainable, and Clinically Not So Relevant

Wolfgang C. Winkelmayer and Georg Heinze†*

Clin J Am Soc Nephrol 8: 520–522, 2013. doi: 10.2215/CJN.02200213

Crucial centerpieces of clinical practice in nephrology are grounded in experience rather than established by hard evidence. One important question that has vexed the community for decades is the comparative effectiveness between hemodialysis (HD) and peritoneal dialysis (PD). In fact, the sole nephrology-specific item among the 100 Initial National Priorities for Comparative Effectiveness Research compiled by the Institute of Medicine's Committee on Comparative Effectiveness Research Prioritization was to "Compare the effectiveness (including survival, hospitalization, quality of life, and costs) of renal replacement therapies (e.g., daily home hemodialysis, intermittent home hemodialysis, conventional in-center dialysis, continuous ambulatory peritoneal dialysis, renal transplantation) for patients of different ages, races, and ethnicities." (1)

Past attempts to conduct a clinical trial have failed because of strong patient preferences for one or the other modality upon receiving detailed education about the respective benefits and shortcomings of these options. As a result, well conducted observational studies provide the best information on which we can rely as clinicians.

Although observational studies are susceptible to biases due to the lack of randomization, recent advances in analytical methods have given us additional tools to better assess the comparative effectiveness of competing strategies (2,3). The current issue of the *CJASN* contains a study by Lukowsky *et al.* (4) that used a rather novel and sophisticated analytical technique, marginal structural modeling, on data from the electronic health records from a large national dialysis provider to determine the association between in-center HD versus PD and mortality in patients who recently initiated dialysis treatment. They found that in patients starting dialysis between 2001 and 2004, PD was associated with a 48% lower mortality than HD during up to 2 years of follow-up (hazard ratio, 0.52; 95% confidence interval, 0.34 to 0.80) (4). These findings present a rather stark contrast to the results presented recently by mostly the same authors using data from the U.S. Renal Data System (USRDS), in which they found that patients who initiated HD and PD between 2002 and 2004 had similar mortality (hazard ratio, 1.03; 95% confidence interval, 0.99 to 1.06) (5).

Justifiably, interested readers may scratch their heads upon such a discrepancy, especially because the cohort used by Lukowsky *et al.* constituted an almost perfect subset of the larger United States–wide cohort in the study by Mehrotra *et al* (5). In this editorial, we attempt to educate consumers of this research who may not be particularly knowledgeable about the technical intricacies of modern statistical methods. We will highlight some of the assumptions, strengths, and pitfalls of using certain modern analytical techniques.

Of the 20 or so cohort studies comparing the outcomes of PD versus HD published in the 1980s and 1990s, most used regular Cox regression for survival analysis. The exposure of interest was the modality on a given day from first dialysis (1, 90, 120, or 180). Patients were usually censored at the time of transplantation, and subsequent modality switches were often ignored, in an attempt to evaluate the consequences of a treatment strategy, reflected by the modality used on the date when follow-up began. Some studies also censored at the time of modality switches, with no consideration of the person-time after these events. Most studies found that PD was associated with higher mortality or that mortality did not differ between the modalities. At the time, it was recognized that the relative mortality of patients using PD versus HD differed by important patient subgroups, such as older age, presence of diabetes, and female sex, all of which were associated with worse outcome on PD. Bloembergen *et al.* first established that the relative mortality also differed by dialysis vintage, with the excess mortality from PD increasing with longer time since dialysis initiation (6). Propensity scores were first used in 2002 to adjust for the observed baseline characteristics of patients using PD versus HD for the initiation of maintenance dialysis (7). Subsequently, propensity-matching was used to further enhance internal validity by restricting analyses to pairs of PD and HD patients who had very similar probabilities of PD treatment (8), now a standard component of USRDS annual reports. To accomplish the same goal, rather than propensity-matching, Mehrotra *et al.* used inverse probability of treatment weighting in a more recent analysis of data from the USRDS (5). This analysis also constituted an important step forward in that they also accounted for differences in informative censoring caused by kidney transplantation, using inverse probability of censoring weighting. The Mehrotra analysis represents the best estimate to date on the

*Division of Nephrology, Stanford University School of Medicine, Palo Alto, California, and †Center for Medical Statistics, Informatics and Intelligent Systems, Section for Clinical Biometrics, Medical University of Vienna, Vienna, Austria

Correspondence: Dr. Wolfgang C. Winkelmayer, Division of Nephrology, Stanford University School of Medicine, 1070 Arastradero Road, Suite 313, Palo Alto, CA 94304. Email: wcw1@stanford.edu

comparative downstream consequences, measured in terms of survival, of a treatment strategy to initiate dialysis on PD versus one that uses HD. Although limited by the exclusion of the first 90 days of dialysis treatment when important (and informative) modality switches occur, the finding of no mortality difference between PD and HD in an intention-to-treat approach that considers all downstream modality switches as (inevitable) consequences of the original modality choice is the best available information for clinicians and patients.

How is it then possible that a study using more or less the same patients from the same era derives such drastically different results, of essentially half the mortality on PD versus HD? Fundamentally, the question asked by Mehrotra *et al.* (5) is different from the one examined in this issue of *CJASN*: Lukowsky *et al.* (4) essentially tried to ask how mortality would have differed between otherwise identical patients ("counterfactual") who had only used PD (over 2 years) compared with otherwise similar patients who had only used HD. Thus, the former approach in Mehrotra *et al.* ignores all treatment switches and considers them inherent consequences of the original modality choice, whereas an analysis embracing all modality switches attempts to exploit all information to determine the "maximum attainable benefit" of one treatment over the other, what we would call the "comparative effectiveness frontier."

Both approaches have one important problem in common, which is controlling for confounding by indication at baseline, arguably a considerable feat given that we know patients using PD are systematically younger, healthier, socioeconomically less challenged, and otherwise more "with it" in less quantifiable ways than patients using HD. This problem is being solved using the same technique in both studies, inverse probability of treatment weighting, a method that effectively eliminates any baseline differences in observed patient characteristics. This enabled Mehrotra *et al.* to analyze their cohort in a manner analogous to an intention-to-treat analysis of a randomized trial. The key assumption is that no residual confounding by unobserved or unobservable, or by observed but inaccurately measured characteristics, is present. Although this is a rather big assumption, it is one that can be understood and reasonably considered by most readers and quantitatively examined using external information or assumptions about possible unmeasured confounders (9).

Ludowsky *et al.* then took their analyses to the next level, by considering and quantitatively embracing modality switches in establishing time-dependent treatment exposure. By doing so, they are crossing the Rubicon and find themselves confronting two fierce (and related) drivers of selection bias: "healthy user" and "sick stopper" phenomena, both treacherous threats to validity. Of particular concern are so-called time-dependent confounders, conditions that are potential consequences of treatment, but at the same time potential determinants of subsequent treatment. For example, in the present study the risk of peritonitis is clearly determined by the previous use of PD as well as a strong determinant of a lasting switch to HD (*i.e.*, nonuse of PD) in the subsequent time period. Thus, episodes and history of peritonitis and other important post-baseline drivers of modality switches need to be taken into account in attempting to tease out the true ("causal") association between modality and mortality.

Marginal structural models are members of a larger family of methods often referred to as "causal methods," a term that has proven to be an irresistible temptation to some researchers who often overinterpreted and oversold the results from these models with little balance or constraint in their discussions. Although causal methods have the proven theoretical property to estimate effects of causal direction and magnitude, they come with inherent assumptions that are usually unreasonable to attain: In addition to absence of any residual confounding at baseline, all time-dependent confounders need to be recorded accurately and in exceedingly discriminate temporal granularity for these models to live up to their theoretical capabilities.

In the present example, all predictors of modality switches would need to be reliably captured over time and explicitly modeled, which is clearly not the case because—other than a handful of laboratories—not a single clinical event or comorbidity was ascertained in a time-dependent way. In the absence of such information, healthy users (PD initiators) and sick stoppers (sick patients depleting the PD and "enriching" the HD cohort over time) rule and drive estimated associations that are bizarrely different from studies, such as that by Mehrotra and colleagues, that analyzed the modality-mortality association in a simpler but more robust intention-to-treat approach.

For the clinician, the most relevant information could come from a well conducted randomized trial whose intention-to-treat analysis would reflect the realities of all downstream consequences of randomized treatment. In the absence of such a trial, observational comparative effectiveness studies of typical patients in typical care settings can provide valuable evidence, with optimal control of confounding by indication at baseline and adequate handling of informative censoring both being key. Although the geek in us remains intrigued by the theoretical properties and possibilities of causal methods, we feel that more needs to be learned about their appropriate use and their requirements in terms of necessary data elements and infrastructure. For the joint decision-making between nephrologist and patient nearing the requirement for long-term dialysis, however, even a perfect estimate of the "comparative effectiveness frontier" would be of little value because it cannot entirely reflect the realities that patients choosing between these modalities encounter.

Acknowledgments

This work was supported by European Commission grant HEALTH-F2-2009-241544, on which Dr. Heinze is an investigator and Dr. Winkelmayer serves as an advisor.

Disclosures

W.C.W. has served as an advisor to Amgen, Bayer, and GlaxoSmithKline.

References

1. Institute of Medicine: *Initial National Priorities for Comparative Effectiveness Research*, Washington, DC, The National Academies Press, 2009
2. Chang TI, Winkelmayer WC: Comparative effectiveness research: What is it and why do we need it in nephrology? *Nephrol Dial Transplant* 27: 2156–2161, 2012
3. Hlatky MA, Winkelmayer WC, Setoguchi S: Epidemiologic and statistical methods for comparative effectiveness research. *Heart Fail Clin* 9: 29–36, 2013

4. Lukowsky LR, Mehrotra R, Kheifets L, Arah OA, Nissenson AR, Kalantar-Zadeh K: Comparing mortality of peritoneal and hemodialysis patients in the first 2 years of dialysis therapy: A marginal structural model analysis. *Clin J Am Soc Nephrol* 8: 619–628, 2013
5. Mehrotra R, Chiu YW, Kalantar-Zadeh K, Bargman J, Vonesh E: Similar outcomes with hemodialysis and peritoneal dialysis in patients with end-stage renal disease. *Arch Intern Med* 171: 110–118, 2011
6. Bloembergen WE, Port FK, Mauger EA, Wolfe RA: A comparison of mortality between patients treated with hemodialysis and peritoneal dialysis. *J Am Soc Nephrol* 6: 177–183, 1995
7. Winkelmayer WC, Glynn RJ, Mittleman MA, Levin R, Pliskin JS, Avorn J: Comparing mortality of elderly patients on hemodialysis versus peritoneal dialysis: A propensity score approach. *J Am Soc Nephrol* 13: 2353–2362, 2002
8. Weinhandl ED, Foley RN, Gilbertson DT, Arneson TJ, Snyder JJ, Collins AJ: Propensity-matched mortality comparison of incident hemodialysis and peritoneal dialysis patients. *J Am Soc Nephrol* 21: 499–506, 2010
9. Schneeweiss S: Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. *Pharmacoepidemiol Drug Saf* 15: 291–303, 2006

Published online ahead of print. Publication date available at www.cjasn.org.

See related article, "Comparing Mortality of Peritoneal and Hemodialysis Patients in the First 2 Years of Dialysis Therapy: A Marginal Structural Model Analysis," on pages 619–628.