

***Robust and Honest Confidence Intervals for Causal Effects:  
Application of a Unified Theory of Parametric, Semi and Nonparametric Statistics  
Based On Higher Dimensional Influence Functions***

James Robins, Professor of Epidemiology and Biostatistics  
Harvard School of Public Health

(the talk is based on joint work with Aad van der Vaart, Eric Tchetgen, and Lingling Li)

**Abstract**

Suppose given data on a dichotomous treatment  $T$ , a response  $Y$ , and a high (say 50) dimensional vector of pre-treatment covariates  $X$ , one wishes to estimate the effect of  $T$  on  $Y$  and is willing to assume no unmeasured confounding (ignorability given  $X$ ). Common analytic approaches include (i) fitting a ‘working’ outcome regression (OR) model for the regression of  $Y$  on  $T$  and  $X$ , and (ii) subclassification by , matching on, or inverse probability of treatment weighting by, an estimate of the propensity score based on a working model for the probability of treatment given  $X$ . Both approaches estimate the treatment effect at the usual parametric rate of square root  $n$  if their respective working model is correct. Here  $n$  is the sample size. . However, approach (i) is biased if the working outcome regression model is misspecified while approach (ii) is biased if the working propensity model is misspecified. A much improved approach is to use a doubly robust estimator that is guaranteed to estimate the treatment effect at the usual parametric rate if either (but not necessarily both ) of the two working models are correct.

However even a doubly robust estimator has the following problem. We get square root  $n$  rates and valid CI of radius  $1/\text{square root } n$  if either working model is correct but an inconsistent estimate and invalid confidence intervals if both models are wrong. Since with high dimensional  $X$ , due to lack of power, we cannot use the data to determine whether even fairly large working models are sufficiently close to being correct that confounding is controlled , it seems a much more honest assessment of our uncertainty to use confidence intervals that (i) shrink to zero (with increasing sample size) at much slower rates than the usual parametric rate but (ii) are much more robust to misspecification of the working models, in the sense that even if the models are not quite correct, the now larger confidence intervals still include the true treatment effect at their nominal coverage rate.

In this talk I describe how this more honest approach can be implemented by using estimators of and confidence intervals for the treatment effect that are higher dimensional  $U$ -statistics. These estimators are derived using a new unified theory of parametric, semi , and nonparametric statistics based on higher order scores (i.e., derivatives of the likelihood), and higher order influence functions that applies equally to both the square-root- $n$  and non-square-root  $n$  problems, reproduces the results previously obtained by the modern theory of non-parametric inference, produces many new non-root-  $n$  results, and most importantly opens up the ability to perform optimal non-root  $n$  inference in complex high dimensional models .