Title

Discussion of Identification, Estimation and Approximation of Risk under Interventions that Depend on the Natural Value of Treatment Using Observational Data, by Jessica Young, Miguel Hernán, and James Robins

Journal

Journal of Causal Inference, 3(1)

ISSN

2193-3677

Authors

van der Laan, Mark J
Luedtke, Alexander R
Díaz, Iván

Publication Date

2014-11-01

DOI

10.1515/em-2014-0012

Peer reviewed

HHS Public Access

Author manuscript

J Causal Inference. Author manuscript; available in PMC 2015 December 01.

Published in final edited form as:

J Causal Inference. 2014 November ; 3(1): 21–31. doi:10.1515/em-2014-0012.

Discussion of Identification, Estimation and Approximation of Risk under Interventions that Depend on the Natural Value of Treatment Using Observational Data, by Jessica Young, Miguel Hernán, and James Robins

Mark J. van der Laan[*],
Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA

Alexander R. Luedtke, and
Division of Biostatistics, University of California, Berkeley, Berkeley, CA, USA, aluedtke@berkeley.edu

Iván Díaz
Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA, idiaz@jhu.edu

Abstract

Young, Hernán, and Robins consider the mean outcome under a dynamic intervention that may rely on the natural value of treatment. They first identify this value with a statistical target parameter, and then show that this statistical target parameter can also be identified with a causal parameter which gives the mean outcome under a stochastic intervention. The authors then describe estimation strategies for these quantities. Here we augment the authors' insightful discussion by sharing our experiences in situations where two causal questions lead to the same statistical estimand, or the newer problem that arises in the study of data adaptive parameters, where two statistical estimands can lead to the same estimation problem. Given a statistical estimation problem, we encourage others to always use a robust estimation framework where the data generating distribution truly belongs to the statistical model. We close with a discussion of a framework which has these properties.

Keywords

dynamic intervention; stochastic intervention; causal inference; targeted learning; semi-parametric model

1 Basic summary of article to set stage for discussion

The authors of this excellent article discuss the identification and estimation of the mean outcome under a dynamic intervention that assigns treatment not only in response to the observed past before treatment but also in response to the actual observed treatment itself

[*]**Corresponding author:Mark J. van der Laan,**laan@berkeley.edu.

under this intervention, where the latter is called the natural value of treatment. We want to congratulate the authors for this nice, welcome, and inspiring article.

Let us consider a single time point example of the type studied in Díaz and van der Laan (2012) in order to provide a very basic summary of the article and set the stage for this discussion. Even though the article considers much more complex, general longitudinal data structures, we believe this simpler example is useful as a starting point for discussion.

Consider a nonparametric structural equation model [NPSEM, Pearl, 2009] with $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Y = f_Y(W, A, U_Y)$, defined by unspecified functions f_W, f_A, f_Y, and some model on the probability distribution of $U = (U_W, U_A, U_Y)$. This defines the model on the full data (U, W, A, Y) and the observed data $O = (W, A, Y)$. Here W are baseline characteristics, A is the intervention node (e.g. treatment variable, missingness indicator, etc.), and Y is the outcome of interest. The full data model (i.e. the allowed set of probability distributions of (U, O)) implies the observed data model (i.e. the allowed set of probability distributions of O). The latter is called the statistical model. Consider now an intervention defined by $W = f_W(U_W)$, $A = f_A(W, U_A)$, $A_d = d(A, W)$, $Y_d = f_Y(W, A_d, U_Y)$, where d is a deterministic function mapping the observed treatment A and covariates W into the treatment value that is assigned to the unit under the intervention. The authors refer to such an intervention as a dynamic intervention that depends on the natural value of treatment. The authors show that the mean outcome under this intervention d is equivalent to the mean outcome under a stochastic intervention g_0^* on A that is only a function of W. The counterfactuals under this stochastic intervention are denoted as $Y_{g_0^*}$. We note that in the special case of a single time point, the natural value of treatment actually equals the observed treatment, while in a longitudinal data structure the natural value of treatment is the counterfactual treatment that would have been observed given the intervention was followed in the past.

Instead of using f^{obs} for the treatment/censoring mechanism, we will use the commonly used notation g. This differs from the use of g in the main text and appendix B of the work of interest where g was, respectively, used to represent dynamic regimes which do not depend on the natural value of treatment and dynamic regimes which may depend on the natural value of treatment. We instead use d to represent a dynamic treatment that may depend on the natural value of treatment. In the main text, the authors use f^d to describe the distribution of such a (possibly stochastic) rule, whereas in this commentary we will focus on deterministic rules d for simplicity. Finally, we use g^* instead of f^{int} for the stochastic intervention that corresponds with the dynamic intervention that relies on the natural value of treatment.

Let $O = (W, A, Y) \sim P_0$ and $O_d = (W, A, A_d, Y_d) \sim P_{0,d}$ where $P_{0,d}$ is called the post-intervention probability distribution. The notation for the variables has changed slightly from the original work to emphasize that we are considering the simpler point treatment case in this commentary. Note that $P_{0,d}$ is determined by the probability distribution of the full data (U, O). The first goal is to identify the full-data parameter $E^{P0,d} Y_d$ as a mapping depending only on the observed data distribution P_0, so that the mean outcome under intervention d can be learned from the observed data. Robins et al. (2004) proposed the extended g-computation formula for this parameter for general longitudinal data structures,

and Richardson and Robins (2013) establish the desired identifiability conditions, under the statistical assumption that the extended *g*-computation formula is well-defined (i.e. the positivity condition holds).

As the authors nicely demonstrate, in many applications, this type of intervention might be considered unrealistic and thereby not interesting: i.e. after the unit has received its natural treatment, one cannot turn around the clock and undo this treatment by replacing it by a new treatment value $d(A, W)$. In these applications, the authors propose an approximation of the target intervention by a dynamic intervention $(A_1, W) \rightarrow d(A_1, W)$, where now A_1 is an *intended* treatment value A_1, instead of the actual realized treatment value, under the assumption that one actually observes such an intended treatment value.

On the other hand, one can also imagine applications in which $(A, W) \rightarrow d(A, W)$ corresponds with an augmentation of the observed treatment value, in which case such an intervention measures the effect of augmenting the treatment by a certain amount that possibly depends on the characteristics of the unit. Thus clearly identification of the mean outcome under such a type of intervention is of both theoretical and practical interest. The SWIG causal graph theory developed in Richardson and Robins (2013) provides a graphical methodology to establish such identification results for general complex, longitudinal data structures. In the single time point example, it is also possible to establish the desired identifiability mathematically, as in Díaz and van der Laan (2012). If

(i) Randomization: A is independent of U_Y, given W, and

(ii) Positivity: $P(A = a|W = w) = 0$ implies $P(d(A, W) = a|W = w) = 0$ for all w in the support of W,

then

$$\begin{aligned}
E_0 Y_d &= E_0 f_Y\left(W, d\left(A, W\right), U_Y\right) \\
&= \sum_{a,w} E_0\left[f_Y\left(W, d\left(A, W\right), U_Y\right) | d\left(A, W\right) = a, W = w\right] P_0\left(d\left(A, W\right) = a, W = w\right) \\
&= \sum_{a,w} E_0\left[f_Y\left(w, a, U_Y\right) | A \epsilon d_w^{-1}\left(a\right), W = w\right] P_0\left(d\left(A, W\right) = a, W = w\right) \\
\text{by (i)} \quad &= \sum_{a,w} E_0\left[f_Y\left(w, a, U_Y\right) | W = w\right] P_0\left(d\left(A, W\right) = a | W = w\right) P_0\left(W = w\right) \\
\text{by (i) and } (ii) \quad &= \sum_{a,w} E_0\left[f_Y\left(w, a, U_Y\right) | A = a, W = w\right] P_0\left(d\left(A, W\right) = a | W = w\right) P_0\left(W = w\right) \\
&= \sum_{a,w} E_{P_0}\left(Y | A = a, W = w\right) g_0^*\left(a | w\right) P_0\left(W = w\right),
\end{aligned}$$

where $d_w^{-1}\left(a\right) = \left\{a' : d_W\left(a', W\right) = a\right\}$ and $g_0^*\left(\cdot | w\right)$ is the conditional distribution of $d(A, W)$ given $W = w$, which can be identified as a function of $P_0(A|W)$.

Note that, in this simple case where such mathematical derivation is tractable, the necessary identification conditions arise naturally in the derivation process. The mathematical derivation above also shows that $E_0 Y_d$ equals the mean outcome $E_0 Y_{g_0^*}$ under a stochastic intervention that replaces the equation $A = f_A(W, U_A)$ by drawing A given W, from g_0^*. This

point is made in general by the authors: the extended *g*-computation formula of the mean outcome under dynamic interventions depending on the natural value of treatment equals the regular *g*-computation formula for the mean outcome under a stochastic intervention that first involves drawing from the actual conditional distribution of treatment, before evaluating the deterministic rule. The authors stress that the equivalence of the two *g*-computation formulas shows that the positivity assumption for this stochastic intervention equals the positivity assumption for the dynamic intervention depending on the natural value of treatment, and that one can use estimators developed for stochastic interventions to estimate the mean under these dynamic interventions depending on natural value of treatment. The authors propose a particular inverse probability of treatment weighted estimator, and contrast estimators based on parametric models and estimators based on semi-parametric statistical models.

2 Discussion items

We focus the discussion on the following points, which are indirectly or directly raised by this article:

Separation of statistical estimation and causal modeling

By recognizing that the identifiability results for two different causal parameters result in the same estimand and statistical model, one can borrow statistical methods and their properties (including their statistical assumptions such as the positivity assumption) developed for one causal parameter to solve the statistical estimation problem for the other.

Enhancing statistical interpretation by using multiple nested identifiability results

Consider two identifiability results that correspond with the same estimand and statistical model, but one result relies on weaker or the same causal assumptions as the other (i.e. one set of assumptions is a subset of the other set of assumptions). Should one not use the former in the interpretation of the statistical results?

Data adaptive target parameters

The estimand for many causal effects of interest corresponds with the estimand for the mean outcome under a stochastic intervention that itself needs to be learned from the data: the mean outcome under a dynamic treatment depending on the natural value of treatment represents one such example. Is it not of interest to define data adaptive target parameters/estimands defined by replacing the stochastic intervention by a data dependent fit of this stochastic intervention? What are the implications for statistical inference for such data adaptive target parameters?

Robust statistical inference

When pursuing statistical inference for the causal quantity of interest, what is the scientific rationale (if any) to select statistical methods that rely on parametric assumptions?

We discuss each of these points in some detail in the remainder of this commentary.

3 Separating statistical estimation from causal modeling

The full data model and full-data target parameter play an important role in obtaining knowledge from subject-matter experts about the data generating experiment and determining the full data target parameter that represents the answer to the scientific question of interest. The full-data model M^F should represent *a priori* knowledge about the phenomena under study, and the full-data target parameter $\Psi^F : M^F \to \mathbb{R}$ should provide the answer $\psi_0^F = \Psi^F\left(P_0^F\right)$ to the scientific question of interest.

Subsequently, it is necessary to establish identifiability of the full-data target parameter from the probability distribution of the observed data, under assumptions which might exceed the assumptions coded by the full-data model. Based on these findings, one will need to commit to a statistical model M and a statistical target parameter, $\Psi^F : M^F \to \mathbb{R}$, with the following two main considerations: (1) the statistical model incorporates the realistic assumptions in the full data model M^F, but not the possibly extra *unrealistic* assumptions that were needed for the identifiability result, in order to guarantee that the statistical model contains the true probability distribution of the data (i.e. $P_0 \in M$) and (2) the target parameter defines an estimand $\psi_0 = \Psi(P_0)$ that approximates the full data parameter value ψ_0^F as best as the data allows. In particular, the estimand ψ_0 should equal the full data target parameter value ψ_0^F when the identifiability assumptions hold. At this point, the statistical estimation problem is well defined. The full-data target parameter and underlying full-data model can be completely ignored in the process of developing estimators and corresponding statistical inference for the statistical parameter.

Consider two of these exercises, possibly starting with different full-data models and full-data parameters, but leading to the same statistical model and statistical target parameter, so that the two statistical estimation problems are identical. In this case, it would be most scientifically coherent to have an estimation procedure that depends only on assumptions affecting the statistical model and statistical target parameter. Therefore, it is very good practice to always be explicit in the formulation of the statistical model M and target parameter Ψ so that the scientific community knows what statistical estimation problem has been addressed, which might be relevant for answering other scientific questions of interest as well. The roadmaps for causal inference presented in Rose and van der Laan (2011), Pearl (2009), and Petersen and van der Laan (2014) make each of these steps explicit. The only role of the full-data model, full data target parameter, and the identifiability result in the estimation process is to generate a statistical model and statistical target parameter. The authors of this article have exemplified this insight by borrowing statistical results for mean outcomes under stochastic interventions, since, for a well-defined stochastic intervention, that problem used the same statistical model and statistical target parameter as used the mean outcome under a dynamic treatments depending on the natural value of treatment.

We have used this general insight into our work as well. For example, in Hubbard et al. (2011), we noted that the identifiability result for a particular type of natural direct effect yielded the same estimand and statistical model as used for the causal effect among the treated, even though the two problems assumed a different time ordering of the data and thus

incomparable sets of causal assumptions. This allowed us to use the efficient and double robust targeted minimum loss-based estimator developed for the causal effect among the treated (Rose and van der Laan 2011) to efficiently estimate this natural direct effect parameter. Similarly, in Lendle et al. (2013) we borrowed the latter TMLE for the effect among the untreated to efficiently estimate the natural direct effect among the untreated. In section A5 of the appendix in Rose and van der Laan (2011), we discuss this general and useful (although trivial in some sense) point in more detail: statistical theorems are invariant to (e.g. non-testable) assumptions that do not change the statistical model and statistical target parameter, allowing us to use the same theorems across very different applications and causal models.

4 Enhancing interpretation of statistical output by referencing multiple identifiability results

The authors discuss identification and estimation of the mean outcome under a dynamic intervention that depends on the natural value of treatment. Suppose that a data analyst uses the extended *g*-computation formula to define an estimator and also provides a 95% confidence interval under statistical assumptions *S*. The data analyst could now make the following two statements: (1) The confidence interval (as a random interval) contains the statistical estimand with probability 0.95 under the statistical assumptions *S*; (2) The confidence interval contains the mean outcome under the dynamic intervention depending on the natural value of treatment with probability 0.95 under the statistical assumptions *S* and the additional identifiability (causal) assumptions *C*. Statement 1) concerns the pure statistical interpretation of the estimand. Statement 2) concerns a statement about the desired causal quantity, under additional assumptions *C*. As shown by the authors, under the same causal assumptions *C*, the estimand also equals the mean outcome under a corresponding stochastic intervention where *A* is drawn from $g_0^*(\cdot|w)$ conditional on $W = w$. Thus, the data analyst could make a third statement: 3) the confidence interval contains the mean outcome under the stochastic intervention g_0^* with probability 0.95 under the same assumptions *S* and *C*. For a longitudinal rather than point treatment data structure, the causal assumptions for 3) can be a subset of the causal assumptions for 2) so that the causal interpretations that can be applied will also vary with which causal assumptions hold.

In our point treatment example (Díaz and van der Laan, 2012), the application of interest might be one where the dynamic intervention $(A, W) \rightarrow d(A, W)$ cannot be carried out in the real world, but the stochastic intervention represents a perfectly plausible experiment of interest. In that case, the additional statement 3) is important for the interpretation of the statistical output.

Let us consider another example. Suppose $O = (W, A, Z, Y) \sim P_0$ and assume the nonparametric structural equation model $W = f_W(U_W)$, $A = f_A(W, U_A)$, $Z = f_Z(W, A, U_Z)$ and $Y = f_Y(W, A, Z, U_Y)$. Suppose that one is concerned with estimation of the natural direct effect defined as

$$\psi_0^F = E_0 \left[Y\left(1, Z_0\right) - Y\left(0, Z_0\right) \right],$$

where $Y(a, Z_0) = f_Y(W, a, Z_0, U_Y)$ and $Z_0 = f_Z(W, 0, U_Z)$, $a \in \{0, 1\}$. One may now use the following identifiability result from the current literature [e.g. Petersen et al. (2006)]: if C_1) $(A, Z) \perp Y(a, z)|W$, for all values (a, z), C_2) $A \perp Z(a)|W$ for all values a, and C_3) $E_0[Y(1, z) - Y(0, z)|Z(0) = z, W] = E_0[Y(1, z) - Y(0, z)|W]$ for all z, then

$$\begin{aligned} \Psi_0^F &= \int \left\{ \bar{Q}_0\left(1, w, z\right) \bar{Q}_0\left(0, w, z\right) \right\} dP_0\left(Z|A=0, w\right) dP_0\left(w\right) \\ &\equiv \Psi\left(P_0\right). \end{aligned}$$

Here we denoted $E_0[Y|A = a, W = w, Z = z]$ with $\bar{Q}_0(a, w, z)$. The data analyst who just computed a 95% confidence interval for ψ_0 under statistical assumptions S can now make the following statements: (1) the confidence interval contains ψ_0 with probability 0.95 under assumptions S; (2) the confidence interval contains Ψ_0^F with probability 0.95 under assumptions S and the above listed causal assumptions C_1, C_2, C_3. However, many will argue that assumption C_3 is particularly hard to defend. One may now note that the estimand ψ_0 also equals $NDE^* \equiv \left[Y_{g^*_{0,1}} - Y_{g^*_{0,0}} \right]$, where $g^*_{0,a}$ is the stochastic intervention on (A, Z) defined as: $g^*_{0,a}\left(a', z|W\right) = I\left(a' = a\right) P_0\left(Z = z|A = 0, W\right), a \epsilon \{0, 1\}$.

In other words, the full data parameter NDE^* is now defined in terms of the mean outcomes under two stochastic interventions on (A, Z) that deterministically set $A = 1$ or $A = 0$ and draws Z from the conditional distribution of Z, given $A = 0$, W (which equals the conditional distribution of Z_0, given W, by C_2). Since NDE^* equals $\Psi(P_0)$ under the randomization assumptions C_1 and C_2 only, the data analysis can now also state 3) the confidence interval contains ψ_0^{F*} with probability 0.95 under S and C_1, C_2. In this manner, one might still obtain reliable inference for NDE^* while reliable inference for NDE is out of the question, due to the indefensible assumption C_3 (Zheng and van der Laan, 2011). Using this approach, Zheng and van der Laan (2012) obtain an identifiability result for a natural direct effect on a time to event outcome, controlling for a time-dependent intermediate process defined in terms of a mean outcome under a stochastic intervention only differing in a static intervention on treatment, where the identifiability only relies on the sequential randomization assumptions required for identification of the mean outcome under these two stochastic interventions.

5 Statistical inference for data adaptive target parameters such as the mean outcome under a stochastic intervention learned from data

It appears that many causal parameters of interest are defined by a mean outcome under a stochastic intervention that itself needs to be learned from data. Let us denote this causal quantity with $E_0 Y_{g_0^*}$, where g_0^* denotes a stochastic intervention that can be identified as a function of P_0. For example, as argued above, the article under discussion defines a full data parameter whose g-computation formula equals the extended g-computation formula for the

mean outcome under a dynamic treatment that depends on the natural value of treatment. The authors might agree that in some applications, in which the dynamic intervention is impossible to carry out and "intended treatment values" are not available, $E_0Y_{g_0^*}$ might be of more interest than the original dynamic treatment parameter. The discussion in this section is relevant in such cases.

Above, we indicated that natural direct effect parameters inspire such analogue natural direct effect parameters which are now defined in terms of stochastic interventions. The mean outcome $E_0Y_{d_0}$ under an optimal dynamic treatment $d_0 = \arg\min_{d \in D} E_0Y_d$ is another example of interest, where $g_0^*=d_0$ is now deterministic but unknown nonetheless. van der Laan and Petersen (2007) and Robins et al. (2008) recommend defining causal quantities (e.g. working marginal structural models) that correspond with realistic dynamic treatment interventions defined as rules that satisfy the strong positivity assumption, where it is often possible to define such rules in terms of the actual (unknown) treatment mechanism g_0. The mean outcome under such a realistic rule is now $E_0Y_{d_0}$ where d_0 is a dynamic treatment defined in terms of g_0. An example of a realistic rule that would belong to this class for the point treatment data structure $O = (W, A, Y)$ is the dynamic treatment $d_0(W) = I(g_0(1|W) > \delta)$ for some $\delta > 0$ that sets $A = 1$ if there is support, but sets $A = 0$ otherwise.

Suppose that $g_n^*=\hat{g}^*(P_n)$ is an estimator of this unknown stochastic intervention g_0^*, mapping the empirical probability distribution P_n of the observed data sample $O_1, \ldots, O_n$ into a realized estimate of g_0^*. Given a data set, we have this estimate g_n^* in our hand. One can imagine that after we have presented our collaborator with a confidence interval for $E_0Y_{g_0^*}$, he or she might ask, what is g_0^* like? The natural answer is to show the collaborator a plot of our estimate g_n^*. Our collaborator might then also consider the target parameter $E_0Y_{g^*}|_{g^*=g_n^*}$, which would tell us what would happen if the tangible rule g_n^* were actually implemented in the population. This parameter is known, given the data, and thus well-defined. Our collaborator might want statistical inference for this data adaptive target parameter as well: that is, one wants a confidence interval that contains the *random* data adaptive parameter with probability 0.95. In van der Laan et al. (2013) we defined such general data adaptive target parameters and established various theorems for statistical inference. In particular, statistical inference can be developed for such data adaptive parameters under appropriate conditions, including a Donsker class condition and a stabilization condition on g_n^* [see theorem 1 in van der Laan et al. (2013)]. The main message is that one can use the same estimator as developed for $EY_{g_0^*}$, but the influence curve is different since it contains no contribution due to estimation of g_0^*. As a consequence, one often ends up with narrower confidence intervals. In fact, it might be difficult or impossible to develop valid inference for $E_0Y_{g_0^*}$, while statistical inference for the data adaptive target parameter can be simply based on the estimator of $E_0Y_{g^*}$ for a fixed g^*, but setting $g^*=g_n^*$.

For example, in van der Laan (2013), van der Laan and Luedtke (2014b), we developed such estimators and such confidence intervals for the mean outcome under an estimate d_n of the optimal dynamic treatment d_0. In this case, one can show that $E_0Y_{d_n} - E_0Y_{d_0}$ is a second-

order term so that one might assume that it is $O_P(1/\sqrt{n})$. Under that assumption and the assumption that the blip functions are nonzero with probability one (Robins and Rotnitzky, 2014; van der Laan and Luedtke, 2014a), the statistical inference for E_0Y_{d0} and $E_0Y_d|_{d=d_n}$ relied on the same estimator and same confidence intervals. Nonetheless, even in this case, the confidence interval for $E_0Y_{d_n}$ avoids reliance on this assumption $E_0Y_{d_n} - E_0Y_{d_0} = O_P(1/\sqrt{n})$, and one obtains better finite sample performance of the estimator and confidence interval even if this assumption holds.

As another example, consider the point treatment data structure and a realistic rule d_0 that sets $A = 1$ if $g_0(1|W) > \delta > 0$ and sets 0 otherwise. Statistical inference for E_0Y_{d0} is problematic due to the fact that the unknown g_0 appears within an indicator defining the treatment rule. As a consequence, the contribution $E_0Y_{d_n}$ - E_0Y_{d0} obtained by estimating this rule might not behave well, so that, contrary to the optimal dynamic treatment example, it is unreasonable to assume that $E_0Y_{d_n} - E_0Y_{d_0} = O_P(1/\sqrt{n})$. If one is willing to assume that $w \to I(g_n(1|w) > \delta)$ has a limit in $L_2(P_0)$ as n gets large, then these serious statistical inference problems for E_0Y_{d0} are completely avoided by simply targeting $E_0Y_{d_n}$ where the given rule d_n is now defined by setting $A = 1$ if $g_n(1|W) > \delta > 0$. Few people would claim that the latter is less interesting than the mean outcome under the unknown realistic rule d_0.

In van der Laan and Luedtke (2014b), our estimator d_n is based on a highly data adaptive super-learner of d_0 developed in Luedtke and van der Laan (2014), so that one might be concerned that the Donsker class condition on d_n might be violated theoretically or negatively affect the finite sample coverage of the confidence interval for $E_0Y_{d_n}$. To deal with this challenge, in van der Laan et al. (2013) and van der Laan (2013), van der Laan and Luedtke (2014b) we started a general theory for estimation and inference for data adaptive parameters, such as theorem 2 in van der Laan et al. (2013) that avoids any conditions on the estimator $\hat{g}^*$, beyond convergence to some fixed g^*. First, we defined data adaptive target parameters of the type $\frac{1}{V}\sum_{v=1}^{V} E_0Y_g|_{g=\hat{g}(P^0_{n,v})}$, where $P^0_{n,v}$ is the training sample for split v in a V-fold sample splitting scheme. That is, one uses the vth training sample to generate a v-specific data adaptive target parameter E_0Y_g with $g = \hat{G}\left(P^0_{n,v}\right)$, and the final data adaptive target parameter is defined as the average of these v-specific data adaptive parameters across the V splits. As shown in van der Laan et al. (2013) one can estimate and obtain inference for such a V-fold data adaptive target parameter by estimating each v-specific data adaptive target parameter based on the v-specific complementary sample. However, if estimators of these v-specific target parameters are highly non-linear such an estimator will suffer from large second-order terms. Therefore, in van der Laan (2012), van der Laan and Luedtke (2014b) we developed a cross-validated TMLE in which only the targeting step relies on cross-validation, and as a consequence the actual estimator of this V-fold data adaptive target parameter will have nice theoretical and practical behavior, not negatively affected by the sample splitting. Specifically, in van der Laan and Luedtke (2014b) we present a cross-validated TMLE of $1/V\sum_{v=1}^{V} E_0Y_{\hat{d}(P^0_{n,v})}$, where $P_n \to \hat{d}(P_n)$ is a super-learner of the

optimal dynamic treatment d_0. In this case the CV-TMLE presents a general method for general cross-validated data adaptive parameters.

We refer to van der Laan et al. (2013) for many other motivating examples demonstrating that statistical inference for data adaptive target parameters opens up a wealth of new scientific questions (that one would not know before looking at the data) and corresponding statistical inference, allowing for data mining to generate the parameters and hypotheses of interest (thereby also avoiding massive multiple testing adjustments).

6 Robust statistical inference: lack of scientific rationale to rely on parametric assumptions

As the authors point out, there is absolutely no reason to use the parametric extended *g*-computation formula method to estimate the desired mean outcome under the dynamic intervention depending on the natural value of treatment, especially since researchers also have access to more robust methods in the semi-parametric model literature. In particular, the authors present an inverse probability of treatment weighted type estimator of the desired extended *g*-computation formula estimand. In this section, we will discuss this point in more detail.

The identification results in causal inference aim to rely on minimal assumptions, in particular, these results typically avoid any statistical assumptions (i.e. restrictions on the probability distribution of the data). That is, many of the identifiability results correspond with nonparametric statistical models. All that hard work for the purpose of reliable inference about causal quantities in this part of the causal inference literature goes to waste if one uses estimators that are biased due to relying on parametric assumptions that are known to be false. It is not scientifically sensible to be nonparametric for the sake of identification but parametric for the sake of estimation given that parametric assumptions are made out of convenience. That is exactly what we do when we use, for example, parametric model-based estimators to estimate the estimand defined by the extended *g*-computation formula, or IPTW estimators based on parametric models for the treatment mechanism. Using such a parametric model-based approach for causal inference makes it less relevant to worry about the causal assumptions since one cannot even trust the estimator of the statistical estimand. This makes one wonder whether there is any theoretical scientific argument to use estimation procedures based on arbitrary parametric assumptions.

One argument might be that estimators based on parametric models can be shown to be asymptotically normally distributed. In other words, we have theorems that show that the confidence intervals have the desired coverage asymptotically under the assumption that these parametric assumptions are true. But what is the point of relying on a theorem whose assumptions are known to be false?

In addition, by not enforcing that a statistical model needs to be correctly specified (i.e. contain the true distribution), different statisticians often end up generating different statistically significant output, even when they are addressing identical statistical estimation problems and have equal access to all the statistical information about the data generating

experiment. The problem here is that the choice of statistical model is viewed as an art instead of a choice driven by scientific knowledge, missing the fact this choice heavily affects the choice of target estimand, the corresponding estimator, and its statistical properties. Some data analysts like to quote “all models are wrong, but some are useful” and use it as an argument that we should not worry too much about the model choice. The truth is that as long as the field of applied statistics is driven by arbitrary model choices, we do not satisfy common sense scientific standards.

Important advances have been made in empirical process theory, weak convergence theory (e.g. van der Vaart and Wellner, 1996), efficiency theory for semi-parametric models (e.g. Bickel et al., 1997), general methods for construction of efficient estimators (e.g. Robins and Rotnitzky, 1992; van der Laan and Robins, 2003; van der Laan and Rose, 2012; Hernan and Robins, 2014), providing us with theorems establishing asymptotic consistency, normality, and efficiency of highly data adaptive estimators in large statistical models. Let us use a concrete demonstration of such a type of theorem concerning the estimation of a pathwise differentiable target parameter $\Psi: M \to \mathbb{R}$ with canonical gradient/efficient influence curve $(P, O) \to D^*(P)(O)$ at P. Given this Ψ and $D^*(P)$ one obtains, by definition of pathwise differentiability, that

$$\Psi(P) - \Psi(P_0) = -P_0 D^*(P) + R_2(P, P_0),$$

where $R_2(P, P_0)$ is a second-order difference between P and P_0 that can be explicitly determined for each choice of target parameter Ψ and model M [see for example, van der Laan, 2012 (2014), for a detailed demonstration]. It is assumed that we select the statistical model M so that one feels confident that $P_0 \in M$.

Consider a substitution estimator $\Psi(P_n^*)$, such as a TMLE based on an initial super-learner-based estimator P_n^0 (van der Laan and Dudoit, 2003; van der Vaart et al., 2006; van der Laan et al., 2006, 2007; Polley et al., 2012) that is then updated into a targeted estimator P_n^*. The estimator $\Psi(P_n^*)$ might also be a parametric *g*-computation estimator relying on a parametric model-based MLE P_n^* of P_0. As shown in van der Laan and Rubin (2006) (or many other subsequent articles, including (van der Laan and Rose, 2012)) $\Psi(P_n^*)$ asymptotically normally distributed and efficient for $\Psi(P_0)$ if

(i) $P_n D^*(P_n^*) = O_P(1/\sqrt{n})$,

(ii) $D^*(P_n^*)$ falls in a P_0-Donsker class with probability tending to 1,

(iii) $P_0(D^*(P_n^*) - D^*(P_0))^2 \to$ in probability, and

(iv) $R_2(P_n^*, P_0) = O_P(1/\sqrt{n})$.

If one uses the TMLE, then condition (i) is automatically satisfied. Condition (ii) would be satisfied, for example, if $D^*(P_n^*)$ falls in the class of multivariate real-valued functions with

uniform sectional variation norm bounded by some $M < \infty$ (Gill et al., 1995), a much less stringent assumption from requiring that P_n^* is estimated in a parametric model. In addition, if one uses a CV-TMLE (Zheng and van der Laan, 2011; Rose and van der Laan, 2011), then condition (ii) can be removed. Let us consider such a CV-TMLE so that the only conditions for asymptotic efficiency are the weak asymptotic consistency condition (iii) and condition (iv). Clearly, condition (iv) is the condition to worry about (if (iv) holds one certainly expects (iii) to hold).

If P_n^* is based on a misspecified parametric model, then there is no hope that $R_2(P_n^*, P_0)$ will converge to zero, i.e. (iv) will not hold. To make this crucial condition as realistic as possible we have promoted the use of super-learning, a cross-validated ensemble learner which incorporates the state of the art of machine learning algorithms and possibly a large variety of parametric model-based estimators. The oracle inequality for the super-learner (see above references) teaches us that we make this condition more and more likely to hold by selecting a library of diverse estimators that grow in size polynomial in sample size. That is, there is no trade off such as that we cannot be too data adaptive, but, on the contrary, we have to push the envelope as much as possible to be maximally data adaptive in order to ensure that $R_2(P_n^*, P_0) = O_P(1/\sqrt{n})$. In addition, under this condition (iv), the estimator is asymptotically efficient and thus also asymptotically regular, a nice by-product for reliable confidence intervals.

In order to move our field forward, we need to fully acknowledge these issues and start defining the estimation problem truthfully. In our work, we defined the field Targeted Learning as the subfield of statistics that is concerned with theory, estimation, and statistical inference (i.e. confidence intervals) for target parameters (representing the answer to the actual scientific question of interest) in realistic statistical models (i.e. incorporating actual knowledge). By necessity, Targeted Learning requires integrating the state of the art in data adaptive estimation, beyond incorporation of subject-matter driven estimators and requires targeting the estimation procedure toward the target parameter of interest. Given these estimators, Targeted Learning requires targeting the estimation procedure toward the target parameter of interest. Targeted minimum loss-based estimation (and its variants such as CV-TMLE, C-TMLE), combined with Super-Learning, provides a general template to construct such targeted substitution estimators (van der Laan and Rubin, 2006; van der Laan and Rose, 2012).

An example of this methodology, relevant to the paper under discussion, is the longitudinal TMLE of summary measures of the mean outcome under dynamic interventions (such as defined by working MSM) in Gruber and van der Laan (2012), Petersen et al. (2013). The TMLE for this problem is inspired by important double robust targeted estimators established in earlier work of Bang and Robins (2005). This TMLE is implemented by the R-package ltmle and fully utilizes the important sequential regression representation presented in Bang and Robins (2005). This TMLE is easily extended to TMLE of summary measures of mean outcomes under stochastic interventions. The extended *g*-computation formula corresponds with an estimated stochastic intervention, so that the statistical inference will now also need to take into account that the stochastic intervention was

estimated. On the other hand, if we go after the mean outcome under a data adaptive fit of the desired stochastic intervention, then the statistical inference is identical to treating the fitted stochastic intervention as known. In this manner, by extending the current ltmle R-package to stochastic (and possibly unknown) interventions (instead of only dynamic interventions), this method would now be accessible to many practitioners, thereby allowing data analysts to significantly improve on the parametric extended *g*-computation formula approach and IPTW estimators relying on parametric models for the treatment mechanism.

We again commend the excellent work of the authors. The field needs more important observations such as this one, which allow the straightforward application of previously described identifiability results and robust estimators to new problems. We further advocate the consideration of newly developed data adaptive target parameters, which often similarly allow for the application of existing estimators to interesting new problems.

Acknowledgments

This research was supported by an NIH grant R01 AI074345-06. AL was supported by the Department of Defense (DoD) through the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program.

Research funding: National Institute of Allergy and Infectious Diseases (Grant/Award Number: “R01 AI074345-06”).

References

Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–972. [PubMed: 16401269]

Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, J. Efficient and Adaptive Estimation for Semiparametric Models. Springer-Verlag; New York: 1997.

Díaz I, van der Laan M. Population intervention causal effects based on stochastic interventions. Biometrics. 2012; 680(2):541–549. ISSN 1541–0420. doi:10.1111/j.1541-0420.2011.01685.x. URL http://dx.doi.org/10.1111/j.1541-0420.2011.01685.x.

Gill RD, van der Laan MJ, Wellner JA. Inefficient estimators of the bivariate survival function for three models. Annales de l’Institut Henri Poincaré. 1995; 31:545–597.

Gruber S, van der Laan MJ. Targeted minimum loss based estimator that outperforms a given estimator. The International Journal of Biostatistics. 2012; 80(1) Article 11. doi: 10.1515/1557–4679.1332.

Hernan, M.; Robins, JM. Causal Inference. Chapman & Hall; London: 2014. Progress

Hubbard, AE.; Jewell, NP.; van der Laan, MJ. Targeted Learning, Springer Series in Statistics. Springer; New York: 2011. Direct effects and effect among the treated, Chapter 8; p. 133-145.ISBN 978-1-4419-9781-4

Lendle SD, Subbaraman MS, van der Laan MJ. Identification and efficient estimation of the natural direct effect among the untreated. Biometrics. 2013; 69:301–317. [PubMed: 23409839]

Luedtke, AR.; van der Laan, MJ. Technical Report 326. UC Berkeley: 2014. Super learning of an optimal dynamic treatment rule. 2014. http://biostats.bepress.com/ucbbiostat/paper326, revised for publication in Journal of Causal Inference

Pearl, J. Causality: Models, Reasoning, and Inference. 2nd. Cambridge University Press; New York: 2009.

Petersen, M.; Schwab, J.; Gruber, S.; Blaser, N.; Schomaker, M.; van der Laan, MJ. Targeted minimum loss based estimation of marginal structural working models. Journal of Causal Inference, to appear. 2013. Technical report. http://biostats.bepress.com/ucbbiostat/paper312/

Petersen ML, Sinisi SE, van der Laan MJ. Estimation of direct causal effects. Epidemiology. 2006; 170(3):276–284. [PubMed: 16617276]

Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014; 250(3):418–426. [PubMed: 24713881]

Polley, EC.; Rose, S.; van der Laan, MJ. Super learning. In: van der Laan, MJ.; Rose, S., editors. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer; New York, Dordrecht, Heidelberg, and London: 2012. p. 43-66.

Richardson TS, Robins JM. Single world intervention graphs (swigs): a unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences. 2013:128. University of Washington Series. Working Paper.

Robins, JM.; Hernán, MA.; Siebert, U. Effects of multiple interventions. In: Ezzati, ADM.; Rodgers Lopez, A.; Murray, CJL., editors. Comparative Quantification of Health Risks: Global and Regional Burden of Disease Attributable to Selected Major Risk Factors. World Health Organization; Geneva: 2004. p. 2191-2230.

Robins JM, Orellana L, Rotnitzky A. Estimation and extrapolation of optimal treatment and testing strategies. Statistics in Medicine. 2008; 27:4678–4721. [PubMed: 18646286]

Robins, JM.; Rotnitzky, A. Recovery of information and adjustment for dependent censoring using surrogate markers. In: Jewell, Nicholas P.; Dietz, Klaus; Farewell, Vernon T., editors. AIDS Epidemiology, Methodological Issues. Birkäuser; Basel, Switzerland: 1992.

Robins J, Rotnitzky A. Discussion of dynamic treatment regimes: technical challenges and applications. Electronic Journal of Statistics. 2014; 80(1):1273, 1289. doi: 10.1214/14-EJS908. http://dx.doi.org/10.1214/14-EJS908.

Rose, S.; van der Laan, MJ. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer; New York: 2011.

van der Laan, MJ. Technical Report 302, Division of Biostatistics. University of California, Berkeley, CA: 2012. Statistical inference when using data adaptive estimators of nuisance parameters.

van der Laan, MJ. Technical Report 317. UC Berkeley: 2013. Targeted learning of an optimal dynamic treatment and statistical inference for its mean outcome. http://biostats.bepress.com/ucbbiostat/paper317

van der Laan MJ. Targeted estimation of nuisance parameters to obtain valid statistical inference. International Journal of Biostatistics. 2014 pii:/j/ijb.ahead-of-print/ijb-2012-0038/ijb-2012-0038.xml. doi: 10.1515/ijb-2012-0038.

van der Laan, MJ.; Dudoit, S. Technical report, Division of Biostatistics. University of California, Berkeley: 2003. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. November 2003

van der Laan MJ, Dudoit S, van der Vaart AW. The cross-validated adaptive epsilon-net estimator. Statistics and Decisions. 2006a; 240(3):373–395.

van der Laan, MJ.; Hubbard, AE.; Kherad, S. Technical Report 314. UC Berkeley: 2013. Statistical inference for data adaptive target parameters. 2013. http://biostats.bepress.com/ucbbiostat/paper314, revised for publication in Biometrics

van der Laan, MJ.; Luedtke, AR. Technical Report 329. UC Berkeley: 2014a. Targeted learning of an optimal dynamic treatment, and statistical inference for its mean outcome.

van der Laan, MJ.; Luedtke, AR. Technical Report 325. UC Berkeley: 2014b. Targeted learning of the mean outcome under an optimal dynamic treatment rule. http://biostats.bepress.com/ucbbiostat/paper325, revised for publication in Journal of Causal Inference

van der Laan MJ, Petersen ML. Causal effect models for realistic individualized treatment and intention to treat rules. International Journal of Biostatistics. 2007; 3(1) Article 3.

van der Laan MJ, Polley E, Hubbard A. Super learner. Statistical Applications in Genetics and Molecular Biology. 2007; 60(25) Article 25. ISSN 1544–6115.

van der Laan, MJ.; Robins, JM. Unified Methods for Censored Longitudinal Data and Causality. Springer; New York: 2003.

van der Laan, MJ.; Rose, S. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer; New York: 2012.

van der Laan MJ, Rubin D. Targeted maximum likelihood learning. The International Journal of Biostatistics. 2006; 20(1) Article 11.

van der Vaart AW, Dudoit S, van der Laan MJ. Oracle inequalities for multi-fold cross-validation. Statistics and Decisions. 2006; 240(3):351–371.

van der Vaart, AW.; Wellner, JA. Weak Convergence and Empirical Processes. Springer-Verlag; New York: 1996.

Zheng, W.; van der Laan, MJ. Cross-validated targeted minimum loss based estimation. In: van der Laan, MJ.; Rose, S., editors. Targeted Learning: Causal Inference for Observational and Experimental Studies. New York; Springer: 2011. p. 459-474.

Zheng, W.; van der Laan, MJ. Technical Report 295. UC Berkeley: 2012. Causal mediation in a survival setting with time-dependent mediators. 2012. http://biostats.bepress.com/ucbbiostat/paper295