University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2010 *Paper* 269

Diagnosing and Responding to Violations in the Positivity Assumption

Maya L. Petersen[*] Kristin Porter[†] Susan Gruber[‡]
Yue Wang[**] Mark J. van der Laan[††]

[*]University of California - Berkeley, mayaliv@berkeley.edu

[†]University of California, Berkeley, kristinporter@berkeley.edu

[‡]University of California, Berkeley, sgruber65@yahoo.com

[**]Department of Clinical Information Services, Novartis Pharmaceuticals Corporation, wangyue@gmail.com

[††]University of California - Berkeley, laan@berkeley.edu

http://biostats.bepress.com/ucbbiostat/paper269

Diagnosing and Responding to Violations in the Positivity Assumption

Maya L. Petersen, Kristin Porter, Susan Gruber, Yue Wang, and Mark J. van der Laan

Abstract

The assumption of positivity or experimental treatment assignment requires that observed treatment levels vary within confounder strata. This article discusses the positivity assumption in the context of assessing model and parameter-speci?c identi?ability of causal e?ects. Positivity violations occur when certain subgroups in a sample rarely or never receive some treatments of interest. The resulting sparsity in the data may increase bias with or without an increase in variance and can threaten valid inference. The parametric bootstrap is presented as a tool to assess the severity of such threats and its utility as a diagnostic is explored using simulated data. Several approaches for improving the identi?ability of parameters in the presence of positivity violations are reviewed. Potential responses to data sparsity include restriction of the covariate adjustment set, use of an alternative pro jection function to de?ne the target parameter within a non-parametric marginal structural model, restriction of the sample, and modi?cation of the target intervention. All of these approaches can be understood as trading o? proximity to the initial target of inference for identi?ability; we advocate approaching this tradeo? systematically.

Diagnosing and responding to violations in the positivity assumption.

Maya L. Petersen (1), Kristin E. Porter (1),
Susan Gruber (1), Yue Wang (2), Mark J. van der Laan (1)

October 21, 2010

Correspondence to :
Maya L. Petersen
101 Haviland Hall, University of California, Berkeley, CA 94110-7358
t: 510-642-0563; f: 510.643.5163; email:mayaliv@berkeley.edu

Abstract

The assumption of positivity or experimental treatment assignment requires that observed treatment levels vary within confounder strata. This article discusses the positivity assumption in the context of assessing model and parameter-specific identifiability of causal effects. Positivity violations occur when certain subgroups in a sample rarely or never receive some treatments of interest. The resulting sparsity in the data may increase bias with or without an increase in variance and can threaten valid inference. The parametric bootstrap is presented as a tool to assess the severity of such threats and its utility as a diagnostic is explored using simulated data. Several approaches for improving the identifiability of parameters in the presence of positivity violations are reviewed. Potential responses to data sparsity include restriction of the covariate adjustment set, use of an alternative projection function to define the target parameter within a non-parametric marginal structural model, restriction of the sample, and modification of the target intervention. All of these approaches can be understood as trading off proximity to the initial target of inference for identifiability; we advocate approaching this tradeoff systematically.

Keywords: experimental treatment assignment, positivity, marginal structural model, inverse probability weight, double robust, causal inference, counterfactual, parametric bootstrap, realistic treatment rule, trimming, stabilized weights, truncation

1 Introduction.

Incomplete control of confounding is a well-recognized source of bias in causal effect estimation- measured covariates must be sufficient to control for confounding in order for causal effects to be identified based on observational data. The identifiability of causal effects further requires sufficient variability in treatment or exposure assignment within strata of confounders. The dangers of causal effect estimation in the absence of adequate data support have long been understood.[1] More recent causal inference literature refers to the need for adequate exposure variability within confounder strata as the assumption of positivity or experimental treatment assignment.[2;3;4] While perhaps less well-recognized than confounding bias, violations and near violations of the positivity assumption can increase both the variance and bias of causal effect estimates, and if undiagnosed can seriously threaten the validity of causal inferences.

Positivity violations can arise for two reasons. First, it may be theoretically impossible for individuals with certain covariate values to receive a given exposure of interest. For example, certain patient characteristics may constitute an absolute contraindication to receipt of a particular treatment. The threat to causal inference posed by such structural or theoretical violations of positivity does not improve with increasing sample size. Second, violations or near violations of positivity can arise in finite samples due to chance. This is a particular problem in small samples, but also occurs frequently in moderate to large samples when the treatment is continuous or can take multiple levels, or when the covariate adjustment set is large and/or contains continuous or multi-level covariates. Regardless of the cause, causal effects may be poorly or non-identified when certain subgroups in a finite sample do not receive some of the treatment levels of interest. In this paper, we will use the term "sparsity" to refer to positivity violations and near-violations arising from either of these causes, recognizing that other types of sparsity can also threaten valid inference.

In this article, we discuss the positivity assumption within a general framework for assessing the identifiability of causal effects. The causal model and target causal parameter are defined using a non-parametric structural equation model (NPSEM) and the positivity assumption is introduced as a key assumption needed for parameter identifiability. The counterfactual or potential outcome framework is then used to review estimation of the target parameter, assessment of the extent to which data sparsity threatens valid inference for this parameter, and practical approaches for responding to such threats. For clarity, we focus on a simple data structure in which treatment is assigned at a single time point. Concluding remarks generalize to more complex longitudinal data structures.

Data sparsity can increase both the bias and variance of a causal effect estimator; the extent to which each are impacted will depend on the estimator used. An estimator-specific diagnostic tool is thus needed to quantify the extent to which positivity violations threaten the validity of inference for a given causal effect parameter (for a given model, data-generating distribution, and finite sample). Wang, et. al. proposed such

a diagnostic based on the parametric bootstrap.[5] Application of a candidate estimator to bootstrapped data sampled from the estimated data generating distribution provides information about the estimator's behavior under a data generating distribution that is based on the observed data. The true parameter value in the bootstrap data is known and can be used to assess estimator bias. A large bias estimate can alert the analyst to the presence of a parameter that is poorly identified, an important warning in settings where data sparsity may not be reflected in the variance of the causal effect estimate.

Once bias due to violations in positivity have been diagnosed, the question remains how best to proceed with estimation. We review several approaches. Identifiability can be improved by extrapolating based on subgroups in which sufficient treatment variability does exist; however, such an approach requires additional parametric model assumptions. Alternative approaches for responding to sparsity include the following: restriction of the sample to those subjects for whom the positivity assumption is not violated (known as trimming); re-definition of the causal effect of interest as the effect of only those treatments that do not result in positivity violations (estimation of the effects of "realistic" or "intention to treat" dynamic regimes); restriction of the covariate adjustment set to exclude those covariates responsible for positivity violations; and, when the target parameter is defined using a marginal structural working model, use of a projection function that focuses estimation on areas of the data with greater support.

As we discuss, all of these approaches change the parameter being estimated by trading proximity to the original target of inference for improved identifiability. We advocate incorporation of this tradeoff into the effect estimator itself. This requires defining a family of parameters, the members of which vary in their proximity to the initial target and in their identifiability. An estimator can then be defined that selects among the members of this family according to some pre-specifed criteria.

1.1 Outline.

The article is structured as follows. Section 2 introduces a non-parametric structural equation model for a simple point treatment data structure, defines the target causal parameter using a non-parametric marginal structural model, and discusses conditions for parameter identifiability with an emphasis on the positivity assumption. Section 3 reviews three classes of causal effect estimators and discusses the behavior of these estimators in the presence of positivity violations. Section 4 reviews approaches for assessing threats to inference arising from positivity violations, with a focus on the parametric bootstrap. Section 5 investigates the performance of the parametric bootstrap as a diagnostic tool using simulated data. Section 6 then applies the diagnostic tool to a real data example. Section 7 reviews methods for responding to positivity violations once they have been diagnosed, and integrates these methods into a general approach to sparsity that is based on defining a family of parameters. Section 8 offers

some concluding remarks and advocates a systematic approach to possible violations in positivity.

2 Framework for causal effect estimation.

We proceed from the basic premise that model assumptions should honestly reflect investigator knowledge. The non-parametric structural equation model (NPSEM) framework of Pearl provides a systematic approach for translating background knowledge into a causal model and corresponding statistical model, defining a target causal parameter, and assessing the identifiability of that parameter.[6] We illustrate this approach using a simple point treatment data structure. We minimize notation by focusing on discrete-valued random variables.

2.1 Model.

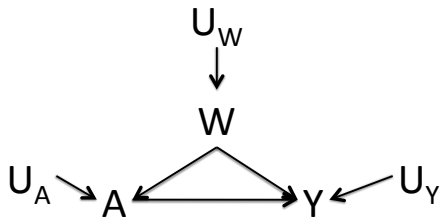
Let W denote a set of baseline covariates on a subject, let A denote a treatment or exposure variable, and let Y denote an outcome. Specify the following structural equation model (with random input $U \sim P_U$):

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y), \end{aligned} \tag{1}$$

where $U = (U_W, U_A, U_Y)$ denotes the set of background factors that deterministically assign values to (W, A, Y) according to functions (f_W, f_A, f_Y). Each of the equations in this model is assumed to represent a mechanism that is autonomous, in the sense that changing or intervening on the equation will not affect the remaining equations, and that is functional, in the sense that the equation reflects assumptions about how the observed data were in fact generated by Nature. In addition, each of the equations is non-parametric, in the sense that its specification does not require assumptions regarding the true functional form of the underlying causal relationships. However, if aspects of the functional form of any of these equations are known based on background knowledge, such knowledge can be incorporated into the model.

A causal graph is derived from a non-parameteric structural equation model by connecting each observed variable to its "parents" (the subset of covariates found in the right hand side of the corresponding structural equation) with arrows emanating from the parents. The causal graph corresponding to Model (1) is given in Figure 1. The background factors U are assumed to be jointly independent in this particular model; or in other words, the model is assumed to be Markov.[6] This assumption is encoded in the absence of double headed arrows between the elements of U in Figure 1. The NPSEM framework can also be applied to non-Markov models.

Figure 1: Causal Graph for Non-Parametric Structural Equation Model (1).



Let the observed data consist of n i.i.d. observations $O_1, \ldots, O_n$ of

$$O = (W, A, Y) \sim P_0.$$

Causal model (1) places no restrictions on the allowed distributions for P_0, and thus implies a non-parametric statistical model.

2.2 Target causal parameter.

A causal effect can be defined in terms of the joint distribution of the observed data under an intervention on one or more of the structural equations, or equivalently, under an intervention on the causal graph. For example, consider the post-intervention distribution of Y under an intervention on the structural model to set $A = a$. Such an intervention corresponds to replacing $A = f_A(W, U_A)$ with $A = a$ in the structural model (1), as follows:

$$\begin{aligned} W &= f_W(U_W) \\ A &= a \\ Y &= f_Y(W, a, U_Y). \end{aligned} \tag{2}$$

The counterfactual outcome that a given subject with background factors u would have had if he or she were to have received treatment level a is denoted $Y_a(u)$.[7;8] This counterfactual can be derived as the solution to the structural equation f_Y in equation system (2) within input $U = u$.

Let F_X denote the distribution of $X = (W, (Y_a : a \in \mathcal{A}))$, where $\mathcal{A}$ denotes the possible values that the treatment variable can take (e.g. $\{0, 1\}$ for a binary treatment). F_X describes the joint distribution of the baseline covariates and counterfactual outcomes under a range of interventions on treatment variable A. A causal effect can be defined

as some function of F_X. For example, a common target parameter for binary A is the average treatment effect

$$E_{F_X}(Y_1 - Y_0), \tag{3}$$

or the difference in expected counterfactual outcome if every subject in the population had received versus had not received treatment.

Alternatively, an investigator may be interested in estimating the average treatment effect separately within certain strata of the population and/or for non-binary treatments. Specification of a marginal structural model (a model on the conditional expectation of the counterfactual outcome given effect modifiers of interest) provides one option for defining the target causal parameter in such cases.[4;9;10] Marginal structural models take the following form:

$$E_{F_X}(Y_a \mid V) = m(a, V \mid \beta), \tag{4}$$

where $V \subset W$ denotes the strata in which one wishes to estimate a conditional causal effect. For example, one might specify the following model:

$$m(a, V \mid \beta) = \beta_1 + \beta_2 a + \beta_3 V + \beta_4 aV.$$

For a binary treatment $\mathcal{A} \in \{0, 1\}$, such a model implies an average treatment effect within stratum $V = v$ equal to $\beta_2 + \beta_4 v$.

The true functional form of $E_{F_X}(Y_a \mid V)$ will generally not be known. One option is to assume that the parametric model $m(a, V \mid \beta)$ is correctly specified, or in other words that $E_{F_X}(Y_a \mid V) = m(a, V \mid \beta)$ for some value β. Such an approach, however, can place additional restrictions on the allowable distributions of the observed data and thus change the statistical model. In order to respect the premise that the statistical model should faithfully reflect the limits of investigator knowledge and not be altered in order to facilitate definition of the target parameter, we advocate an alternative approach in which the target causal parameter is defined using a non-parametric marginal structural model. Under this approach the target parameter β is defined as the projection of the true causal curve $E_{F_X}(Y_a \mid V)$ onto the specified model $m(a, V \mid \beta)$ according to some projection function $h(a, V)$:

$$\beta(F_X, m, h) = \underset{\beta}{\operatorname{argmin}} E_{F_X}\left[\sum_{a \in \mathcal{A}} (Y_a - m(a, V|\beta))^2 h(a, V)\right].^{11} \tag{5}$$

When $h(a, V) = 1$, the target parameter β corresponds to an unweighted projection of the entire causal curve onto the model $m(a, V|\beta)$; alternative choices of h correspond to placing greater emphasis on specific parts of the curve (i.e. on certain (a, V) values).

Use of a non-parametric marginal structural model such as (5) is attractive because it allows the target causal parameter to be defined within the original statistical model. However, this approach by no means absolves the investigator from careful consideration of marginal structural model specification. A poorly specified model

$m(a, V|\beta)$ may result in a target parameter that provides a poor summary of the features of the true causal relationship that are of interest.

In the following sections we discuss the parameter $\beta(F_X, m, 1)$ as the target of inference, corresponding to a focus on estimation of the treatment-specific mean for all levels $a \in \mathcal{A}$ within strata of V as projected onto model m, with projection $h(a, V) = 1$ chosen to reflect a focus on the entire causal curve. To simplify notation we use β to refer to this target parameter unless otherwise noted.

2.3 Identifiability.

We assess whether the target parameter β of the counterfactual data distribution F_X is identified as a parameter of the observed data distribution P_0 under causal Model (1). Because Model (1) is Markov, we have that

$$P_{F_X}(Y_a = y) = \sum_w P_0(Y = y|W = w, A = a)P_0(W = w), \tag{6}$$

identifying the target parameter β according to projection (5).[6] This identifiability result is often referred to as the G-computation formula.[2;3;12] The weaker assumption of randomization, or the assumption that A and Y_a are conditionally independent given W, is also sufficient for identifiability result (6) to hold.

Randomization Assumption:

$$A \coprod Y_a|W \text{ for all } a \in \mathcal{A}.^{2;3;12} \tag{7}$$

Whether or not a given structural model implies that assumption (7) holds can be assessed directly from the graph through the back door criterion.[6]

2.3.1 The need for experimentation in treatment assignment.

The G-computation formula (6) is only a valid formula if the conditional distributions in the formula are well-defined. Let $g_0(a \mid W) \equiv P_0(A = a \mid W), a \in \mathcal{A}$ denote the conditonal distribution of treatment variable A under the observed data distribution P_0. If one or more treatment levels of interest do not occur within some covariate strata, the conditional probability $P_0(Y = y|A = a, W = w)$ will not be well-defined for some value(s) (a, w) and the identifiability result (6) will break down.

A simple example provides intuition into the threat to parameter identifiability posed by sparsity of this nature. Consider an example in which $W = I(woman)$, A is a binary treatment, and no women are treated ($g_0(1|W = 1) = 0$). In this data generating distribution there is no information regarding outcomes among treated women. Thus, as long as there are women in the target population (i.e. $P_0(W = 1) > 0$), the average treatment effect $E_{F_X}(Y_1 - Y_0)$ will not be identified without additional parametric assumptions.

This simple example illustrates that a given causal parameter under a given model may be identified for some joint distributions of the observed data but not for others. An additional assumption is thus needed to ensure identfiability. We begin by presenting the strong version of this assumption, needed for the identification of $P_{F_X}((Y_a = y, W = w) : a, y, w)$ in a non-parametric model.

Strong Positivity Assumption:

$$\inf_{a \in \mathcal{A}} g_0(a \mid W) > 0, \text{- a.e.} \tag{8}$$

The strong positivity assumption, or assumption of experimental treatment assignment (ETA), states that each possible treatment level occurs with some positive probability within each strata of W.

Parametric model assumptions may allow the positivity assumption to be weakened. In the example above, an assumption that the treatment effect is the same among treated men and women would result in identification of the average treatment effect (3) based on extrapolation from the estimated treatment effect among men (assuming that other identifiability assumptions were met). Parametric model assumptions of this nature are particularly dangerous, however, because they extrapolate to regions of the joint distribution of (A, W) that are not supported by the data. Such assumptions should be approached with caution and adopted only when they have a solid foundation in background knowledge.

In addition to being model-specific, the form of the positivity assumption needed for identifiability is parameter-specific. Many target causal parameters require much weaker versions of positivity than (8). To take one simple example, if the target parameter is $E(Y_1)$, the identifiability result only requires that $g_0(1|W) > 0$ hold; it doesn't matter if there are some strata of the population in which no one was treated. Similarly, the identifiability of $\beta(F_X, m, h)$, defined using a marginal structural model, relies on a weaker positivity assumption.

Positivity Assumption for $\beta(F_X, h, m)$:

$$\sup_{a \in \mathcal{A}} \frac{h(a, V)}{g(a|W)} < \infty, \text{- a.e.} \tag{9}$$

Choice of projection function $h(a, V)$ used to define the target parameter thus has implications for how strong an assumption on positivity is needed for identifiability. In Section 7 we consider specification of alternative target parameters that allow for weaker positivity assumptions than (8), including parameters indexed by alternative choices of $h(a, V)$. For now we focus on the target parameter β indexed by the choice $h(a, V) = 1$ and note that (8) and (9) are equivalent for this parameter.

Once a target parameter has been specified, an assessment of its identifiability should precede estimation. Causal graphs provide a tool for assessment of identifiability assumption (7); however, an additional tool is needed to assess threats to identifiability arising from positivity violations or near violations. Section 4 reviews approaches

for diagnosing such threats, with a focus on the parametric bootstrap. Because the impact of positivity violations is estimator-specific, we first review several common estimators of β and discuss their behavior in the face of sparsity.

3 Estimator-specific behavior in the face of positivity violations.

Let $\Psi(P_0)$ denote the target parameter of the observed data distribution, which under the assumptions of randomization (7) and positivity (9) equals the target causal parameter $\beta(F_X, m, h)$. Estimators of this parameter are denoted $\hat{\Psi}(P_n)$, where P_n is the empirical distribution of a sample of n i.i.d observations from P_0. We use $Q_{0W}(w) \equiv P_0(W = w)$, $Q_{0Y}(y|A, W) \equiv P_0(Y = y|A, W)$, and $\bar{Q}_0(A, W) \equiv E_0(Y|A, W)$. Recall that $g_0(a|W) \equiv P_0(A = a|W)$. We review three classes of estimators $\hat{\Psi}(P_n)$ of β that employ estimators of distinct parts of the observed data likelihood. Maximum likelihood-based substitution estimators (also referred to as "G-computation" estimators) employ estimators of $Q_0 \equiv (Q_{0W}, \bar{Q}_0)$. Inverse probability weighted estimators employ estimators of g_0. Double robust estimators employ estimators of both g_0 and Q_0. A summary of these estimators is provided in Table 1. Their behavior in the face of positivity violations is illustrated in Section 5 and previous work.[11;13;14;15;16]

We focus our discussion on bias in the point estimate of the target parameter β. While estimates of the variance of β can also be biased when data are sparse, methods exist to improve variance estimation. The non-parametric bootstrap provides one straightforward approach to variance estimation in setting where the central limit theorem may not apply as a result of sparsity; alternative approaches to correct for biased variance estimates are also possible.[17] These methods will not, however, protect against misleading inference if the point estimate itself is biased.

3.1 The G-computation estimator.

The G-computation estimator $\hat{\Psi}_{Gcomp}(P_n)$ provides a mapping from the empirical data distribution P_n to a parameter estimate $\hat{\beta}_{Gcomp}$. $\hat{\Psi}_{Gcomp}(P_n)$ is a substitution estimator based on identifiability result (6). It is implemented based on an estimator of $Q_0 \equiv (Q_{0W}, \bar{Q}_0)$ and its consistency relies on the consistency of this estimator.[2;3] Q_{0W} can generally be estimated based on the empirical distribution of W. However, even when positivity is not violated, the dimension of A, W is frequently too large for $\bar{Q}_0$ to be estimated simply by evaluating the mean of Y within strata of (A, W). Due to the curse of dimensionality, estimation of $\bar{Q}_0$ under a non-parametric or semi-parametric statistical model thus frequently requires data-adaptive approaches such as cross-validated loss-based learning.[18;19;20]

Given an estimator $\bar{Q}_n$ of $\bar{Q}_0$, the G-computation estimator can be implemented by generating a predicted counterfactual outcome for each subject under each possible

Table 1: Overview of three classes of causal effect estimator.

G-computation Estimator (Section 3.1)	
Needed for Implementation:	Estimator Q_n of Q_0
Needed for Consistency:	Q_n is a consistent estimator of Q_0
Response to Sparsity:	Extrapolates based on Q_n Sparsity can amplify bias due to model misspecification
IPTW Estimator (Section 3.2.)	
Needed for Implementation:	Estimator g_n of g_0
Needed for Consistency:	g_n is a consistent estimator of g_0 g_0 satisfies positivity
Response to Sparsity:	Does not extrapolate based on Q_n Sensitive to positivity violations and near violations
DR Estimators (Section 3.3.)	
Needed for Implementation:	Estimator g_n of g_0 and Q_n of Q_0
Needed for Consistency:	g_n is consistent or Q_n is consistent g_n converges to a distribution that satisfies positivity
Response to Sparsity:	Can extrapolate based on Q_n Without positivity, relies on consistency of Q_n

treatment: $\hat{Y}_{a,i} = \bar{Q}_n(a, W_i)$ for $a \in \mathcal{A}$, $i = 1, ..., n$. The estimate $\hat{\beta}_{Gcomp}$ is then obtained by regressing $\hat{Y}_a$ on a and V according to the model $m(a, V \mid \beta)$, with weights based on the projection function $h(a, V)$.

When all treatment levels of interest are not represented within all covariate strata (i.e. assumption (8) is violated), some of the conditional probabilities in the non-parametric G-computation formula (6) will not be defined. A given estimate $\bar{Q}_n$ may allow the G-computation estimator to extrapolate based on covariate strata in which sufficient experimentation in treatment level does exist. Importantly, however, this extrapolation depends heavily on the model for $\bar{Q}_0$ and the resulting effect estimates will be biased if the model used to estimate $\bar{Q}_0$ is misspecified.

Moore et. al. illustrate the bias that can arise in the G-computation estimator when simple model fitting algorithms such as forward and backward selection are used to estimate $\bar{Q}_0(A, W)$.[15] While more sophisticated model fitting techniques can improve estimator performance, they do not resolve the potential for data sparsity to result in substatial bias. One possible source of positivity violations is collinearity between a confounder or set of confounders and the treatment or exposure of interest. If data-adaptive methods are used to fit $\bar{Q}(A, W)$, covariates that are collinear or highly correlated with treatment may be dropped from a model in which treatment is forced. If these covariates are also confounders, resulting effect estimates will be biased.

Traditional Multivariable Approaches. A traditional approach to effect estimation in many fields is to estimate $\bar{Q}_0 \equiv E_0(Y|A, W)$ using a multivariable regression

model and to report the estimated coefficient on A (or some transformation of this coefficient, such as its exponentiated value) as the estimated causal effect. In some cases such an estimate is equivalent to the G-computation estimate. For example, if the target of inference is the average treatment effect for binary A, a traditional analysis might fit the model $\hat{E}(Y|A,W) = \hat{\beta}_0 + \hat{\beta}_1 A + k(W)$ and report an effect estimate of $\hat{\beta}_1$. In this case, $\hat{\beta}_1$ will be equivalent to $\hat{\beta}_{Gcomp}$ (assuming the same model is used for $\bar{Q}_n$ when implementing the G-computation estimator).

In many cases, however, the coefficient on A in the multivariable regression model used to estimate $\bar{Q}_0$ represents a distinct estimand. For example, for binary Y a common approach is to fit a logistic regression model such as $\hat{E}(Y|A,W) = 1/(1 + \exp^{-(\hat{\beta}_0 + \hat{\beta}_1 A + k(W))})$. Here $\exp(\hat{\beta}_1)$, which is commonly reported as the causal effect estimate of interest, is an estimate of the conditional odds ratio and is not equivalent to either the average treatment effect or the marginal odds ratio. If G-computation is used to estimate either of the latter two quantities then clearly the resulting estimates will not be equivalent. Traditional regression approaches can consistently estimate causal parameters when identifiability conditions are met and $\bar{Q}_n$ is correctly specified; however, care must be taken to ensure that the parameter estimated corresponds to the causal question of interest.

3.2 The Inverse Probability of Treatment Weighted estimator

The IPTW estimator $\hat{\Psi}_{IPTW}(P_n)$ provides a mapping from the empirical data distribution P_n to a parameter estimate $\hat{\beta}_{IPTW}$ based on an estimator g_n of $g_0(A|W)$.[10;21] The estimator is defined as the solution in β to the following estimating equation:

$$0 = \sum_{i=1}^{n} \frac{h(A_i, V_i)}{g_n(A_i \mid W_i)} \frac{d}{d\beta}(m(A_i, V_i|\beta))(Y - m(A_i, V_i \mid \beta)), \tag{10}$$

where $h(A,V)$ is the projection function used to define the target causal parameter $\beta(F_X, m, h)$ according to (5). The IPTW estimator of β can be implemented as the solution to a weighted regression of the outcome Y on treatment A and effect modifiers V according to model $m(A,V|\beta)$, with weights equal to $\frac{h(A,V)}{g_n(A|W)}$. Consistency of $\hat{\Psi}_{IPTW}(P_n)$ requires that g_0 satisfies positivity and that g_n is a consistent estimator of g_0. As with $\bar{Q}_0$, g_0 can be estimated using loss-based learning and cross validation. Depending on choice of projection function, implementation may further require estimation of $h(A,V)$; however, the consistency of the IPTW estimator does not depend on consistent estimation of $h(A,V)$.

The IPTW estimator is particularly sensitive to bias due to data sparsity. Bias can arise due to structural positivity violations (positivity may not hold for g_0) or may occur because by chance certain covariate and treatment combinations are not represented or sparsely represented in a given finite sample. In the latter case, $g_n(a|W = w)$ will have values of zero or close to zero for some (a, w) even when positivity holds for

g_0 and g_n is consistent.[5;13;14;16;15] As fewer individuals within a given covariate stratum receive a given treatment, the weights of those rare individuals who do receive the treatment become more extreme. The disproportionate reliance of the causal effect estimate on the experience of a few unusual individuals can result in substantial finite sample bias.

While values of $g_n(a \mid W)$ remain positive for all $a \in \mathcal{A}$, elevated weights inflate the variance of the effect estimate and can serve as a warning that the data may poorly support the target parameter. However, as the number of individuals within a covariate stratum who receive a given treatment level shifts from few (each of whom receive a large weight and thus elevate the variance) to none, estimator variance can decrease while bias increases rapidly. In other words, when $g_n(a|W = w) = 0$ for some (a, w), the weight for a subject with $A = a$ and $W = w$ is infinity; however, as no such individuals exist in the dataset, the corresponding threat to valid inference will not be reflected in either the weights or in estimator variance.

Weight truncation. Weights are commonly truncated or bounded in order to improve the performance of the IPTW estimator in face of data sparsity.[5;15;16;22;23] Weights are truncated at either a fixed or relative level (for example, at the 1st and 99th percentiles), thereby reducing the variance arising from large weights and limiting the impact of a few possibly non-representative individuals on the effect estimate. This advantage comes at a cost, however, in the form of increased bias due to misspecification of the treatment model g_n, a bias that does not decrease with increasing sample size. In Section 5, we use simulated data to illustrate the performance of the IPTW estimator under a range of values for weight truncation, illustrate how even in the face of sparsity, weight truncation can increase rather than decrease estimator mean squared error, and discuss how the parametric bootstrap can be used to approach truncation.

Stabilized Weights. Use of projection function $h(a, V) = 1$ implies the use of unstabilized weights. In contrast, stabilized weights, corresponding to a choice of $h(a, V) = g(a|V)$ (where $g(a|V)$ denotes $P_0(A = a|V)$) are generally recommended for the implementation of marginal structural model-based effect estimation. The choice of $h(a, V) = g(a|V)$ results in a weaker positivity assumption, according to (9), by allowing the IPTW estimator to extrapolate to sparse areas of the joint distribution of (A, V) using the model $m(a, V|\beta)$. For example, if A is an ordinal variable with multiple levels, $V = \{\}$, and the target parameter is defined using the model $m(a, V|\beta) = \beta_0 + \beta_1 a$, the IPTW estimator with stabilized weights will extrapolate to levels of A that are sparsely represented in the data by assuming a linear relationship between Y_a and $a \in \mathcal{A}$. We note, however, that when the target parameter β is defined using a non-parametric marginal structural model according to (5) (an approach that acknowledges that the model $m(A, V|\beta)$ may be misspecified), the use of stabilized versus unstabilized weights corresponds to a shift in the target parameter via choice of an alternative projection function.[11]

3.3 Double Robust estimators.

Double robust (DR) approaches to estimation of β include the augmented inverse probability weighted estimator (A-IPTW) and targeted maximum likelihood estimator (TMLE) (which for the target parameter $\beta(F_X, h, m)$ corresponds to the extended double robust parametric regression estimator of Sharfstein *et. al.*).[4;24;25;26;27;28] Implementation of the double robust estimators requires estimators of both Q_0 and g_0; as with the IPTW and G-computation estimators, a non-parametric loss-based approach can be employed in the estimation of both. An implementation of the TMLE estimator of the average treatment effect $E(Y_1 - Y_0)$ for binary A is available in the R package `tmleLite`; an implementation of the A-IPTW estimator in the point treatment setting is available in the R package `cvDSA` (both available at http://www.stat.berkeley.edu/ laan/Software/index.html). Prior literature provides further details regarding implementation and theoretical properties.[4;11;13;24;26;27;28]

Double robust estimators remain consistent if either 1) g_n is a consistent estimator of g_0 and g_0 satisfies positivity; or, 2) Q_n is a consistent estimator of Q_0 and g_n converges to a distribution g^* that satisfies positivity. Thus when positivity holds, these estimators are truly double robust, in the sense that consistent estimation of either g_0 or Q_0 results in a consistent estimator. When positivity fails, however, the consistency of the double robust estimators relies entirely on consistent estimation of Q_0. In the setting of positivity violations, double robust estimators are thus faced with the same vulnerabilities as the G-computation estimator.

In addition to illustrating how positivity violations increase the vulnerability of double robust estimators to bias resulting from inconsistent estimation of Q_0, these asymptotic results have practical implications for the implementation of the double robust estimators. Specifically, they suggest that use of an estimator g_n that satisfies positivity (or in other words, that yields predicted values in $[0 + \gamma, 1 - \gamma]$ where γ is some small number) can improve finite sample performance. One way to achieve such bounds is by truncating the predicted probabilities generated by g_n, similar to the process of weight truncation described for the IPTW estimator.

Alternative double robust estimators are available that make more sophisticated choices in estimating g_0. In particular, the collaborative targeted maximum likelihood estimator (C-TMLE) selects an estimator g_n aimed at optimizing estimation of the target parameter as assessed by the targeted log likelihood. In particular this implies that the C-TMLE estimator includes in the fit of g_n only those covariates that improve estimation of the target.[29] However, when the target parameter is poorly identified due to positivity violations, C-TMLE may be forced to accept significant bias in its aim to optimize mean squared error for the target parameter. Diagnostic procedures remain essential to alert the analyst that such a tradeoff is occurring.

4 Diagnosing bias due to positivity violations.

Positivity violations can result in substantial bias, with or without a corresponding increase in variance, regardless of the causal effect estimator used. Practical methods are thus needed to diagnose and quantify estimator-specific positivity bias for a given model, parameter and sample. Cole and Hernan suggest a range of informal diagnostic approaches when the IPTW estimator is applied.[16] Basic descriptive analyses of treatment variability within covariate strata can be helpful; however, this approach quickly becomes unwieldy when the covariate set is moderately large and includes continuous or multi-level variables. Examination of the distribution of the estimated weights can also provide useful information as near violations of the positivity assumption will be reflected in large weights. As noted by these authors and discussed above, however, well-behaved weights are not sufficient in themselves to ensure the absence of positivity violations.

An alternative formulation is to examine the distribution of the estimated propensity score values given by $g_n(a|W)$ for $a \in \mathcal{A}$. Values of $g_n(a|W)$ close to 0 for any a constitute a warning regarding the presence of positivity violations. We note that examination of the propensity score distribution is a general approach not restricted to the IPTW estimator. However, while useful in diagnosing the presence of positivity violations, examination of the estimated propensity scores does not provide any quantitative estimate of the degree to which such violations are resulting in estimator bias and may pose a threat to inference. The parametric bootstrap can be used to provide an optimistic bias estimate specifically targeted at bias caused by positivity violations and near-violations.[5]

4.1 The parametric bootstrap as a diagnostic tool.

We focus on the bias of estimators that target a parameter of the observed data distribution; this target observed data parameter is equal under the randomization assumption (7) to the target causal parameter. (Divergence between the target observed data parameter and target causal parameter when (7) fails is a distinct issue not addressed by the proposed diagnostic.) The bias in an estimator is the difference between the true value of the target parameter of the observed data distribution and the expectation of the estimator applied to a finite sample from that distribution:

$$Bias(\hat{\Psi}, P_0, n) = E_{P_0}\hat{\Psi}(P_n) - \Psi(P_0),$$

where we recall that $\Psi(P_0)$ is the target observed data parameter, $\hat{\Psi}(P_n)$ is an estimator of that parameter (which may be a function of g_n, Q_n or both), and P_n denotes the empirical distribution of a sample of n i.i.d observations from the true observed data distribution P_0.

Bias in an estimator can arise due to a range of causes. First, the estimators g_n and/or Q_n may be inconsistent. Second, g_0 may not satisfy the positivity assumption. Third,

consistent estimators g_n and/or Q_n may still have substantial finite sample bias. This latter type of finite sample bias arises in particular due to the curse of dimensionality in a non-parametric or semi-parametric model when g_n and/or Q_n are data-adaptive estimators, although it can also be substantial for parametric estimators. Fourth, estimated values of g_n may be equal or close to zero or one, despite use of a consistent estimator g_n and a distribution g_0 that satisfies positivity. The relative contribution of each of these sources of bias will depend on the model, the true data generating distribution, the causal effect estimator, and the finite sample.

The parametric bootstrap provides a tool that allows the analyst to explore the extent to which bias due to any of these causes is affecting a given parameter estimate. The parametric bootstrap-based bias estimate is defined as:

$$\widehat{Bias}_{PB}(\hat{\Psi}, \hat{P}_0, n) = E_{\hat{P}_0}\hat{\Psi}(P_n^{\#}) - \Psi(\hat{P}_0), \tag{11}$$

where $\hat{P}_0$ is an estimate of P_0 and $P_n^{\#}$ is the empirical distribution of a bootstrap sample obtained by sampling from $\hat{P}_0$. In other words, the parametric bootstrap is used to sample from an estimate of the true data generating distribution, resulting in multiple simulated data sets. The true data generating distribution and target parameter value in the bootstrapped data are known. A candidate estimator is then applied to each bootstrapped data set and the mean of the resulting estimates compared with the known "truth" (i.e. the true parameter value for the bootstrap data generating distribution).

We focus on a particular algorithm for parametric bootstrap-based bias estimation, which specifically targets the component of estimator-specific finite sample bias due to violations and near violations of the positivity assumption. The goal is not to provide an accurate estimate of bias, but rather to provide a diagnostic tool that can serve as a "red flag" warning that positivity bias may pose a threat to inference. The distinguishing characteristic of the diagnostic algorithm is its use of an estimated data generating distribution $\hat{P}_0$ that both approximates the true P_0 as closely as possible and is compatible with the estimators $\bar{Q}_n$ and/or g_n used in $\hat{\Psi}(P_n)$. In other words, $\hat{P}_0$ is chosen such that the estimator $\hat{\Psi}$ applied to bootstrap samples from $\hat{P}_0$ is guaranteed to be consistent unless g_0 fails to satisfy the positivity assumption or g_n is truncated. As a result, the parametric bootstrap provides an optimistic estimate of finite sample bias, in which bias due to model misspecification other than truncation is eliminated.

We refer informally to the resulting bias estimate as $ETA.Bias$ because in many settings it will be predominantly composed of bias from the following sources: 1) violation of the positivity assumption by g_0; 2) truncation, if any, of g_n in response to positivity violations; and, 3) finite sample bias arising from values of g_n close to zero or one (sometime referred to as practical violations of the positivity assumption). The term $ETA.Bias$ is imprecise because the bias estimated by the proposed algorithm will also capture some of the bias in $\hat{\Psi}(P_n)$ due to finite sample bias of the estimators g_n and $\bar{Q}_n$ (a form of sparsity only partially related to positivity). Due to the curse of dimensionality, the contribution of this latter source of bias may be substantial when

g_n and/or Q_n are data-adaptive estimators in a non-parametric or semi-parametric model. However, the proposed diagnostic algorithm will only capture a portion of this bias because, unlike P_0, $\hat{P}_0$ is guaranteed to have a functional form that can be well-approximated by the data-adaptive algorithms employed by g_n and Q_n.

The diagnostic algorithm for $ETA.Bias$ is implemented as follows.

Step 1. Estimate P_0. Estimation of P_0 requires estimation of Q_{0W}, g_0, and Q_{0Y}, (i.e. estimation of $P_0(W = w)$, $P_0(A = a|W = w)$, and $P_0(Y = y|A = a, W = w)$ for all (w, a, y)). We define $Q_{\hat{P}_0 W} = Q_{P_n W}$ (or in other words, use an estimate based on the empirical distribution of the data), $g_{\hat{P}_0} = g_n$, and $\bar{Q}_{\hat{P}_0} = \bar{Q}_n$. Note that the estimators $Q_{P_n W}$, g_n, and $\bar{Q}_n$ were all needed for implementation of the IPTW, G-compuation, and DR estimators; the same estimators can be used here. Additional steps may be required to estimate the entire conditional distribution of Y given (A, W) (beyond the estimate of its mean given by $\bar{Q}_n$). The true target parameter for the known distribution $\hat{P}_0$ is only a function of $Q_n = (Q_{P_n W}, \bar{Q}_n)$, and $\Psi(\hat{P}_0)$ is the same as the G-computation estimator (using Q_n) applied to the observed data:

$$\Psi(\hat{P}_0) = \hat{\Psi}_{Gcomp}(P_n).$$

Step 2. Generate $P_n^{\#}$ by sampling from $\hat{P}_0$. In the second step, we assume that $\hat{P}_0$ is the true data generating distribution. Bootstrap samples $P_n^{\#}$, each with n i.i.d observations, are generated by sampling from $\hat{P}_0$. For example, W can be sampled from the empirical, a binary A can be generated as a Bernoulli with probability $g_n(1|W)$, and a continuous Y can be generated by adding a $N(0, 1)$ error to $\bar{Q}_n(A, W)$ (alternative approaches are also possible).

Step 3. Estimate $E_{\hat{P}_0}\hat{\Psi}(P_n^{\#})$. Finally, the estimator $\hat{\Psi}$ is applied to each bootstrap sample. Depending on the estimator being evaluated, this step involves applying the estimators g_n, Q_n or both to each bootstrap sample. If Q_n and/or g_n are data-adaptive estimators, the corresponding data-adaptive algorithm should be rerun in each bootstrap sample; otherwise, the coefficients of the corresponding models should be refit. $ETA.Bias$ is calculated by comparing the mean of the estimator $\hat{\Psi}$ across bootstrap samples ($E_{\hat{P}_0}\hat{\Psi}_{IPTW}(P_n^{\#})$) with the true value of the target parameter under the bootstrap data generating distribution ($\Psi(\hat{P}_0)$).

The parametric bootstrap-based diagnostic applied to the IPTW estimator is available as an R function `check.ETA` in the `cvDSA` package.[5] The routine takes the original data as input and performs bootstrap simulations under user-specified information such as functional forms for $m(a, V \mid \beta)$, g_n and Q_n. Application of the bootstrap to the IPTW estimator offers one particularly sensitive assessment of positivity bias because, unlike the G-computation and double robust estimators, the IPTW estimator can not extrapolate based on $\bar{Q}_n$. However, this approach can be applied to any causal effect estimator, including estimators introduced in Section 7 that trade off identifiability for proximity to the target parameter. In assessing the threat posed by positivity violations the bootstrap should ideally be applied to both the IPTW estimator and the estimator of choice.

Remarks on interpretation of the bias estimate. We caution against using the parametric bootstrap for any form of bias correction. The true bias of the estimator is $E_{P_0}\hat{\Psi}(P_n) - \Psi(P_0)$, while the parametric bootstrap estimates $E_{\hat{P}_0}\hat{\Psi}(P_n^{\#}) - \Psi(\hat{P}_0)$. The performance of the diagnostic thus depends on the extent to which $\hat{P}_0$ approximates the true data generating distribution. This suggests the importance of using flexible data-adaptive algorithms to estimate P_0. Regardless of estimation approach, however, when the target parameter $\Psi(P_0)$ is poorly identified due to positivity violations $\Psi(\hat{P}_0)$ may be a poor estimate of $\Psi(P_0)$. In such cases one would not expect the parametric bootstrap to provide a good estimate of the true bias. Further, the $ETA.Bias$ implementation of the parametric bootstrap provides a deliberately optimistic bias estimate by excluding bias due to model misspecifcation for the estimators g_n and $\bar{Q}_n$.

Rather, the parametric bootstrap is proposed as a diagnostic tool. Even when the data generating distribution is not estimated consistently, the bias estimate provided by the parametric bootstrap remains interpretable in the world where the estimated data generating mechanism represents the truth. If the estimated bias is large, an analyst who disregards the implied caution is relying on an unsubstantiated hope that first, he or she has inconsistently estimated the data generating distribution but still done a reasonable job estimating the causal effect of interest; and second, the true data generating distribution is less affected by positivity (and other finite sample) bias than is the analyst's best estimate of it.

The threshold level of $ETA.Bias$ that is considered problematic will vary depending on the scientific question and the point and variance estimates of the causal effect. With that caveat, we suggest the following two general situations in which $ETA.Bias$ can be considered a "red flag" warning: 1) when $ETA.Bias$ is of the same magnitude as (or larger than) the estimated standard error of the estimator; and, 2) when the interpretation of a bias-corrected confidence interval would differ meaningfully from initial conclusions.

Use of a data-adaptive algorithm for $\bar{Q}_n$ may result in exclusion of those elements of W responsible for positivity violations. Bootstrap data sampled from the resulting estimate $\hat{P}_0$ will contain less sparsity than is present in the true data generating distribution, resulting in an underestimate of bias due to positivity violations. One approach to improving the sensitivity of the diagnostic in such settings is to force the estimator $\bar{Q}_n(A, W)$ to include all W known or thought to contribute to positivity violations. The estimated propensity score provides a convenient dimension reduction of exactly those W. Thus a more comprehensive approach to identifying threats to inference due to positivity bias could involve implementing the bootstrap-based $ETA.Bias$ diagnostic using several estimators Q_n, including an estimator that forces inclusion of A but allows W to be selected data adaptively and an estimator that forces inclusion of both A and the propensity score but allows W to be selected data-adaptively. Finally, when the targeted maximum likelihood estimator is implemented, the bootstrap can sample from the targeted estimate of the likelihood it provides, an estimate in which Q_n is already a function of the propensity score. We demon-

strate the propensity score-based approaches in Section 5; however, the performance of the diagnostic when data-adaptive approaches are used and positivity violations are present, as well as the relative performance of various approaches to improving diagnostic performance in such settings, should be investigated further.

5 Simulations.

Data were simulated under three data generating distributions with different degrees and sources of positivity violations. In each set of simulations, four estimators described in Section 3, G- computation, IPTW, A-IPTW, and TMLE, were applied. (Specifically, TMLE was implemented with a logistic fluctuation for continuous and binary Y.)[30] For each simulation, each estimator was implemented using a range of approaches to estimate g_0 and Q_0. Both the behavior of the estimator and the performance of the parametric bootstrap as a diagnostic tool were investigated under each scenario. The objectives of these simulations were (1) to demonstrate how different estimators are affected differently by violations of the positivity assumption; (2) to demonstrate the value and limitations of the bootstrap-based diagnostic in different settings; and (3) to illustrate how the diagnostic might be used in practice to inform interpretation of results. We provide selected simulation results here; additional results together with simulation code are available at http://www.stat.berkeley.edu/ laan/Software/index.html.

5.1 Data generating distributions.

All three simulations used a binary A, and targeted the same causal parameter, $E(Y_1 - Y_0)$ or the average treatment effect. This target parameter is a special case of $\beta(F_X, m, h)$ corresponding to $V = \{\}$ and use of marginal structural model $m(a|\beta) = \beta_0 + \beta_1 a$, and a case in which G-computation corresponds to traditional regression-based adjustment. The true target parameter value $\Psi(P_0) = \beta_1$.

The simulations were based, to varying degrees, on a data generating distribution used by Freedman and Berk.[31] Two baseline covariates, $W = (W_1, W_2)$, were generated bivariate normal, $N(\mu, \Sigma)$, with $\mu_1 = 0.5$, $\mu_2 = 1$, and $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$. The true conditional expectation of Y, given A and W, $\bar{Q}_0(A, W) \equiv E_0(Y|A, W)$ is given by:

$$\bar{Q}_0(A, W) = 1 + A + W_1 + 2W_2,$$

and Y was generated as $\bar{Q}_0(A, W) + U$, with $U \sim N(0, 1)$. The true value of the target parameter $\Psi(P_0) = 1$. The true treatment mechanism, $g_0(1|W) \equiv P_0(A = 1|W)$ is given by:

$$g_0(1|W) = \Phi(0.5 + 0.25W_1 + 0.75W_2),$$

where Φ is the CDF of the standard normal distribution. In other words, the treatment mechanism, or conditional probability of treatment given covariates, was based on a probit model.

Simulation 1: For our first simulation, we modified g_0 to reduce the extent of positivity violations by multiplying all coefficients in g_0 by 0.3. Therefore, the true treatment mechanism in Simulation 1 is given by:

$$g_0(1|W) = \Phi(0.3(0.5 + 0.25W_1 + 0.75W_2)).$$

With this treatment mechanism, $g_0 \in [0.48, 0.92]$. We generated 250 samples of size 1000 for this simulation.

Simulation 2: Simulation 2 is identical to Freedman and Berk's original simulation described above. Again we generated 250 samples of size 1000. In this simulation, $g_0 \in [0.001, 1]$.

Simulation 3: For this simulation, $W1 \sim N(0.5, 1)$ and $W2 \sim Bernoulli(0.5)$. We varied $\bar{Q}_0(A, W)$ such that:

$$\bar{Q}_0(1, W) = expit(-1 + 5A + W_1 + 10W_2).$$

Binary Y was generated as a Bernoulli trial with probability $\bar{Q}_0(1, W)$. The target parameter $E(Y_1 - Y_0)$ for binary Y corresponds to the risk difference. For this simulation, $\Psi(P_0) = 0.29$. The treatment mechanism for this simulation is given by:

$$g_0(1|W) = expit(-3 - 1W_1 + 9W_2).$$

Binary A was generated as a Bernoulli trial with probability $g_0(1|W)$. Under this treatment mechanism A and $W2$ are collinear, with correlation 0.95 and $g_0 \in [0.001, 1]$. For this simulation, we generated 250 samples of size 200 instead of size 1000. The smaller sample size increased the sparsity in the data.

5.2 Investigation of estimator behavior and the performance of the parametric bootstrap-based diagnostic.

The bias, variance, and mean squared error of each estimator were estimated by applying the estimator to 250 samples drawn from the three data generating distributions above. For Simulations 1 and 2, each of the four estimators was implemented with each of the following three approaches: 1) use of a correctly specified model to estimate both $\bar{Q}_0$ and g_0 (a specification referred to as "*Qcgc*"); 2) use of a correctly specified model to estimate $\bar{Q}_0$ but omission of W_2 from the model used to estimate g_0 ("*Qcgm*"); and, 3) omission of W_2 from $\bar{Q}_n$ while correctly specifying the model

used to estimate g_0 ("$Qmgc$"). In Simulation 3, each of the four estimators was implemented using correctly specified models for both g_0 and $\bar{Q}_0$ ($Qcgc$), and using forward stepwise selection based on AIC to estimate both $\bar{Q}_0$ and g_0, using the R function `step` and forcing A to be included in $\bar{Q}_n$ ("$Qdgd1$") . The double robust and IPTW estimators were further implemented using the following sets of bounds for the values of g_n: $[0, 1]$ (or no bounding), $[0.025, 0.975]$, $[0.05, 0.95]$, and $[0.1, 0.9]$. For the IPTW estimator, the latter three bounds correspond to truncation of the unstabilized weights at $[1.03, 40]$, $[1.05, 20]$, and $[1.11, 11.1]$.

The parametric bootstrap was then applied to estimate $ETA.Bias$ for 10 of the 250 samples from each of the three simulations. For each sample and for each model specification ($Qcgc$, $Qmgc$ and $Qcgm$ for Simulations 1 and 2; and $Qcgc$ and $Qdgd1$ for Simulation 3), the estimates Q_n and g_n were used to draw 1000 parametric bootstrap samples. Specifically, W was drawn from the empirical distribution for that sample; A was generated as a series of Bernoulli trials with probability $g_n(1|W)$, and Y was generated either by adding a $N(0, 1)$ error to $\bar{Q}_n(A, W)$ (for continuous Y in Simulations 1 and 2) or as a series of Bernoulli trials with probability $\bar{Q}_n(1|A, W)$ (for binary Y in Simulation 3). Each candidate estimator was then applied to each bootstrap sample. In Simulation 3, an alternative implementation of the diagnostic based on including the propensity score in $\bar{Q}_n$ was also applied ("$Qdgd2$"). Specifically, the stepwise algorithm was forced to retain both A and the estimated propensity score $g_n(1|W)$ as covariates in the estimate $\bar{Q}_n$ used to generate the bootstrap samples.

For the specifications $Qcgc$, $Qmgc$ and $Qcgm$, the models used to estimate g_0 and $\bar{Q}_0$ were held fixed across bootstrap samples and their coefficients refit in each bootstrap sample. For the data-adaptive approaches $Qdgd1$ and $Qdgd2$, the stepwise selection algorithm was rerun in each bootstrap sample, and was forced to retain A in $\bar{Q}_n$. $ETA.Bias$ was estimated for each of the 10 samples as the difference between the mean of the bootstrapped estimator and the initial G-computation estimate $\Psi(\hat{P}_0) = \hat{\Psi}_{Gcomp}(P_n)$ in that sample.

5.3 Results: Simulation 1.

In this simulation the positivity assumption is not violated, and as expected, all four estimators performed well when correctly specified models were used to estimate g_0 and $\bar{Q}_0$. The bias, variance, and MSE for each estimator are shown in Table 2. As described in Section 3, misspecification of the model used to estimate $\bar{Q}_0$ introduced bias in the G-computation estimator, misspecification of the model used to estimate g_0 introduced bias in the IPTW estimator, and the double robust estimators remained minimally biased if the model for either $\bar{Q}_0$ or g_0 was correctly specified.

Table 3 reports the mean and variance of the estimated $ETA.Bias$ for each estimator and model specification across 10 of the 250 original samples. Consistent with the results in Table 2, the estimated $ETA.Bias$ was minimal and varied little across the 10 samples. The parametric bootstrap would not have raised a red flag for any of the

Table 2: Performance of estimators by specification in Simulation 1: g_0 in [0.48,0.92], shown for unbounded g_n only. Results are based on 250 samples of size 1000.

	Qcgc			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
G-COMP	1.5e-03	5.9e-03	5.9e-03	1.5e-03	5.9e-03	5.9e-03	2.6e-01	1.9e-02	8.5e-02
IPTW	6.0e-03	9.2e-03	9.2e-03	2.6e-01	2.1e-02	9.0e-02	6.0e-03	9.2e-03	9.2e-03
A-IPTW	2.6e-04	6.2e-03	6.2e-03	5.9e-04	6.0e-03	6.0e-03	7.2e-04	6.7e-03	6.7e-03
TMLE	-6.7e-06	6.2e-03	6.2e-03	3.9e-04	6.0e-03	6.0e-03	5.0e-04	6.6e-03	6.6e-03

estimators in this scenario, an appropriate result given Table 2.

5.4 Results: Simulation 2.

Simulation 2 introduced substantial data sparsity. Table 4 demonstrates the effect of positivity violations and near-violations on estimator behavior across 250 samples. The G-computation estimator remained minimally biased when the estimator $\bar{Q}_n$ was consistent; use of inconsistent $\bar{Q}_n$ resulted in bias. Given consistent estimators $\bar{Q}_n$ and g_n, the IPTW estimator was more biased than the other three estimators, as expected given the practical positivity violations present in the simulation. For this particular data-generating distribution and choice of misspecified model, misspecification of g_n increased the bias of the IPTW estimator further; however, this will not always be the case.

The finite sample performance of the A-IPTW and TMLE estimators was also affected by the presence of practical positivity violations. The DR estimators achieved the lowest MSE when 1) $\bar{Q}_n$ was consistent and 2) g_n was inconsistent but satisfied positivity (as a result either of truncation or of omission of W_2, a major source of positivity bias). Interestingly, in this simulation TMLE still did quite well when $\bar{Q}_n$ was inconsistent and the model used for g_n was correctly specified but its values bounded at $[0.025, 0.925]$.

Choice of bound imposed on g_n affected both the bias and variance of the IPTW, A-IPTW, and TMLE estimators. As expected, truncation of the IPTW weights improved the variance of the estimator but increased bias. Without additional diagnostic information, an analyst who observed the dramatic decline in the variance of the IPTW estimator that occurred with weight truncation might have concluded that truncation improved estimator performance; however, in this simulation weight truncation increased MSE. In contrast, and as predicted by theory, use of bounded values of g_n decreased MSE of the double robust estimators in spite of the inconsistency introduced to g_n.

Table 5 shows the mean and variance of the estimates of $ETA.Bias$ across 10 of the 250 samples. Based on the results shown in Table 4, a red flag diagnostic for the

Table 3: True finite sample bias by specification (based on 250 samples of sample size 1000 with consistent g_n and Q_n) and mean and variance of estimated $ETA.Bias$ (based on the first 10 of the 250 samples) in Simulation 1: g_0 in [0.48,0.92], shown for unbounded g_n only.

		G-COMP	IPTW	A-IPTW	TMLE
	True finite sample bias	1.51e-03	5.95e-03	2.61e-04	-6.71e-06
Qcgc	Mean(ETA.Bias)	-4.21e-04	5.92e-04	-5.43e-04	-6.94e-04
	Variance(ETA.Bias)	2.23e-06	2.81e-06	2.34e-06	2.35e-06
	Mean(ETA.Bias)/True Bias	-2.79e-01	9.94e-02	-2.08e+00	1.03e+02
Qcgm	Mean(ETA.Bias)	6.17e-04	1.27e-03	4.17e-04	2.42e-04
	Variance(ETA.Bias)	7.32e-06	1.57e-05	6.48e-06	6.54e-06
	Mean(ETA.Bias)/True Bias	4.09e-01	2.14e-01	1.60e+00	-3.61e+01
Qmgc	Mean(ETA.Bias)	6.99e-04	1.51e-03	4.78e-04	3.05e-04
	Variance(ETA.Bias)	6.37e-06	8.18e-06	7.27e-06	7.25e-06
	Mean(ETA.Bias)/True Bias	4.63e-01	2.54e-01	1.83e+00	-4.54e+01

Table 4: Performance of estimators by specification and by bound on g_n in Simulation 2: g_0 in [0.001,1]. Results are based on 250 samples of size 1000.

Bound on g_n	**Qcgc**			**Qcgm**			**Qmgc**		
	Bias	Var	MSE	Bias	Var	MSE	Bias	Var	MSE
G-COMP									
None	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
[0.025,0.975]	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
[0.05,0.95]	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
[0.1,0.9]	0.007	0.009	0.009	0.007	0.009	0.009	1.145	0.025	1.336
IPTW									
None	0.544	0.693	0.989	1.547	0.267	2.660	0.544	0.693	0.989
[0.025,0.975]	1.080	0.090	1.257	1.807	0.077	3.340	1.080	0.090	1.257
[0.05,0.95]	1.437	0.059	2.123	2.062	0.054	4.306	1.437	0.059	2.123
[0.1,0.9]	1.935	0.043	3.787	2.456	0.043	6.076	1.935	0.043	3.787
A-IPTW									
None	0.080	0.966	0.972	-0.003	0.032	0.032	-0.096	16.978	16.987
[0.025,0.975]	0.012	0.017	0.017	0.006	0.017	0.017	0.430	0.035	0.219
[0.05,0.95]	0.011	0.014	0.014	0.009	0.014	0.014	0.556	0.025	0.334
[0.1,0.9]	0.009	0.011	0.011	0.008	0.011	0.011	0.706	0.020	0.519
TMLE									
None	0.251	0.478	0.540	0.026	0.059	0.060	-0.675	0.367	0.824
[0.025,0.975]	0.016	0.028	0.028	0.005	0.021	0.021	-0.004	0.049	0.049
[0.05,0.95]	0.013	0.019	0.020	0.010	0.016	0.017	0.163	0.027	0.054
[0.1,0.9]	0.010	0.014	0.014	0.009	0.013	0.013	0.384	0.018	0.166

presence of bias due to positivity violations was needed for the IPTW estimator at all levels of bounding g_n, and for the TMLE estimator with unbounded g_n. (The A-IPTW estimator had a small to moderate level of bias with unbounded g_n; however the high variance of this estimator would have alerted an analyst to sparsity.) The parametric bootstrap correctly identified the presence of substantial $ETA.Bias$ in the IPTW estimator regardless of truncation level and in the TMLE estimator with unbounded g_n. It suggested minimal $ETA.Bias$ for the remaining estimators.

For correctly specified Q_n and g_n (g_n unbounded), the diagnostic captured 78% and 69% of the true finite sample bias of the IPTW and TMLE estimators, respectively. The fact that the true bias was underestimated in both cases illustrates a key limitation of the parametric bootstrap- its performance suffers when the target estimator is not asymptotically normally distributed.[32] Bounding g_n improved the ability of the bootstrap to accurately diagnose bias by improving estimator behavior (in addition to adding a new source of bias due to use of inconsistent g_n). This finding suggests that practical application of the bootstrap to a given estimator should at minimum generate $ETA.Bias$ estimates for a single low level of bounding g_n in addition to any unbounded estimate. When g_n was bounded, the estimated $ETA.Bias$ for the IPTW estimator captured 96-98% of the true finite sample bias. The $ETA.Bias$ for the TMLE estimator with bounded g_n was accurately estimated to be minimal. As expected, misspecification of g_n or $\bar{Q}_n$ by excluding a key confounder lead to an estimated data generating distribution with less sparsity than the original, and as a result the parametric bootstrap underestimated the true extent of positivity bias for these model specifications.

While use of an unbounded g_n resulted in an underestimate of the true degree of $ETA.Bias$ for the IPTW and TMLE estimators, in this simulation the parametric bootstrap would still have functioned well as a diagnostic in each of the 10 samples considered. Tables 6 and 7 report the output that would have been available to an analyst applying the parametric bootstrap to the IPTW and TMLE estimators for each of the 10 samples. For both unbounded g_n for both estimators, the estimated $ETA.Bias$ was similar in magnitude or larger than the estimated standard error of the estimator, and was of significant magnitude relative to the point estimate of the causal effect. The magnitude of $ETA.Bias$ increased for the IPTW estimator when bounded g_n was used.

Table 5 further demonstrates how the parametric bootstrap can be used to investigate the tradeoffs between bias due to weight truncation/bounding of g_n and positivity bias. The parametric bootstrap accurately diagnosed both an increase in the bias of the IPTW estimator with increasing truncation and a reduction in the bias of the TMLE estimator with truncation. When viewed in light of the standard error estimates under different levels of truncation, the diagnostic would have accurately suggested that truncation of g_n for the TMLE estimator was beneficial, while truncation of the weights for the IPTW estimator was of questionable benefit. (The parametric bootstrap can also be used to provide a more refined approach to choosing an optimal truncation constant based on estimated MSE.[23])

Table 5: True finite sample bias (based on 250 samples of size 1000 with Qcgc) and mean and variance of estimated $ETA.Bias$ (based on the first 10 of the 250 samples) by specification and by bound on g_n in Simulation 2: g_0 in [0.001,1].

		Bound on g_n			
		None	[0.025,0.975]	[0.05,0.95]	[0.1,0.9]
G-COMP	True finite sample bias	7.01e-03	7.01e-03	7.01e-03	7.01e-03
	Mean(ETA.Bias)	-8.51e-04	-8.51e-04	-8.51e-04	-8.51e-04
Qcgc	Variance(ETA.Bias)	5.63e-06	5.63e-06	5.63e-06	5.63e-06
	Mean(ETA.Bias)/True bias	-1.21e-01	-1.21e-01	-1.21e-01	-1.21e-01
	Mean(ETA.Bias)	2.39e-04	2.39e-04	2.39e-04	2.39e-04
Qcgm	Variance(ETA.Bias)	1.37e-05	1.37e-05	1.37e-05	1.37e-05
	Mean(ETA.Bias)/True bias	3.41e-02	3.41e-02	3.41e-02	3.41e-02
	Mean(ETA.Bias)	5.12e-04	5.12e-04	5.12e-04	5.12e-04
Qmgc	Variance(ETA.Bias)	1.22e-05	1.22e-05	1.22e-05	1.22e-05
	Mean(ETA.Bias)/True bias	7.30e-02	7.30e-02	7.30e-02	7.30e-02
IPTW	True finite sample bias	5.44e-01	1.08e+00	1.44e+00	1.93e+00
	Mean(ETA.Bias)	4.22e-01	1.04e+00	1.40e+00	1.90e+00
Qcgc	Variance(ETA.Bias)	9.55e-03	2.19e-02	2.34e-02	2.39e-02
	Mean(ETA.Bias)/True Bias	7.76e-01	9.63e-01	9.73e-01	9.80e-01
	Mean(ETA.Bias)	1.34e-01	4.83e-01	7.84e-01	1.23e+00
Qcgm	Variance(ETA.Bias)	1.96e-03	1.08e-02	1.83e-02	2.40e-02
	Mean(ETA.Bias)/True Bias	2.46e-01	4.48e-01	5.46e-01	6.37e-01
	Mean(ETA.Bias)	2.98e-01	7.39e-01	9.95e-01	1.35e+00
Qmgc	Variance(ETA.Bias)	3.75e-03	9.65e-03	1.09e-02	1.36e-02
	Mean(ETA.Bias)/True Bias	5.48e-01	6.84e-01	6.93e-01	7.00e-01
A-IPTW	True finite sample bias	7.99e-02	1.25e-02	1.07e-02	8.78e-03
	Mean(ETA.Bias)	1.86e-03	2.80e-03	5.89e-05	1.65e-03
Qcgc	Variance(ETA.Bias)	1.51e-04	1.12e-05	4.68e-06	1.51e-05
	Mean(ETA.Bias)/True bias	2.32e-02	2.24e-01	5.50e-03	1.88e-01
	Mean(ETA.Bias)	-3.68e-04	-6.36e-04	2.56e-05	5.72e-04
Qcgm	Variance(ETA.Bias)	7.54e-05	1.16e-05	1.15e-05	1.53e-05
	Mean(ETA.Bias)/True bias	-4.60e-03	-5.09e-02	2.39e-03	6.51e-02
	Mean(ETA.Bias)	-3.59e-04	1.21e-04	-1.18e-04	-1.09e-03
Qmgc	Variance(ETA.Bias)	2.19e-04	1.04e-05	1.41e-05	5.31e-06
	Mean(ETA.Bias)/True bias	-4.50e-03	9.70e-03	-1.10e-02	-1.25e-01
TMLE	True finite sample bias	2.51e-01	1.60e-02	1.31e-02	9.98e-03
	Mean(ETA.Bias)	1.74e-01	4.28e-03	2.65e-04	1.84e-03
Qcgc	Variance(ETA.Bias)	3.26e-03	2.32e-05	6.26e-06	2.23e-05
	Mean(ETA.Bias)/True bias	6.94e-01	2.67e-01	2.02e-02	1.84e-01
	Mean(ETA.Bias)	2.70e-02	-3.07e-04	2.15e-04	7.74e-04
Qcgm	Variance(ETA.Bias)	2.88e-04	1.50e-05	1.27e-05	1.46e-05
	Mean(ETA.Bias)/True bias	1.08e-01	-1.92e-02	1.64e-02	7.76e-02
	Mean(ETA.Bias)	1.11e-01	9.82e-04	-2.17e-04	-1.47e-03
Qmgc	Variance(ETA.Bias)	8.95e-04	2.59e-05	2.52e-05	6.48e-06
	Mean(ETA.Bias)/True bias	4.44e-01	6.13e-02	-1.66e-02	-1.47e-01

Table 6: IPTW estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 2: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$
1	0.207	0.203	0.473	1.462	0.196	1.092	2.119	0.197	1.456	2.815	0.201	1.965
2	1.722	0.197	0.425	2.339	0.192	1.047	2.665	0.190	1.413	3.033	0.192	1.924
3	1.957	0.184	0.306	2.192	0.182	0.876	2.493	0.181	1.217	2.880	0.183	1.717
4	1.926	0.206	0.510	2.648	0.200	1.310	2.973	0.198	1.672	3.311	0.199	2.170
5	2.201	0.192	0.565	2.267	0.193	1.158	2.551	0.196	1.510	3.029	0.202	2.000
6	0.035	0.236	0.520	2.450	0.196	1.146	2.767	0.192	1.504	3.154	0.195	1.999
7	1.799	0.180	0.346	1.999	0.180	0.996	2.433	0.181	1.338	2.915	0.184	1.813
8	-0.471	0.215	0.420	1.938	0.193	1.007	2.400	0.194	1.398	2.978	0.196	1.922
9	2.749	0.184	0.391	2.769	0.185	0.977	2.828	0.186	1.326	3.088	0.189	1.822
10	-0.095	0.228	0.263	1.289	0.210	0.788	1.847	0.206	1.139	2.513	0.201	1.636
Mean	1.203	0.203	0.422	2.135	0.193	1.040	2.508	0.192	1.397	2.972	0.194	1.897

Table 7: TMLE estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 2: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$
1	0.827	0.197	0.172	0.982	0.105	0.001	0.975	0.077	0.001	0.965	0.063	0.003
2	0.734	0.114	0.153	1.094	0.100	-0.001	1.144	0.089	0.003	1.115	0.077	0.003
3	1.379	0.105	0.087	1.171	0.089	0.010	1.167	0.084	0.002	1.136	0.071	0.011
4	0.237	0.089	0.252	0.886	0.077	0.006	0.968	0.071	-0.001	1.016	0.071	-0.003
5	2.548	0.182	0.245	1.205	0.130	0.008	1.095	0.095	0.000	1.035	0.077	0.006
6	0.533	0.228	0.234	1.137	0.122	0.010	1.126	0.084	-0.001	1.083	0.071	0.000
7	1.781	0.184	0.150	1.143	0.138	0.002	1.159	0.095	0.001	1.128	0.077	0.004
8	1.066	0.114	0.188	0.950	0.095	0.003	0.919	0.084	-0.005	0.944	0.071	0.000
9	1.974	0.114	0.161	1.278	0.084	0.007	1.235	0.077	-0.001	1.176	0.071	-0.004
10	0.628	0.173	0.099	0.785	0.1451	-0.004	0.838	0.126	0.004	0.907	0.089	-0.003
Mean	1.171	0.170	0.174	1.063	0.114	0.004	1.063	0.095	0.000	1.051	0.071	0.002

to the analyst having an bias estimate due to misspecification of g_0. It's important to remind the reader that ETA.Bias includes bias both due to ETA and to bounding g_n.

We recommend that the parametric bootstrap be applied to the IPTW estimator in addition to the analyst's estimator of choice. Tables 5 and 6 illustrate the benefits of this approach. Diagnosis of substantial bias in the IPTW estimator due to positivity violations would have alerted an analyst that the G-computation estimator was relying heavily on extrapolation, and that the double robust estimators were sensitive to bias arising from misspecification of the model used to estimate $\bar{Q}_0$.

5.5 Results: Simulation 3.

This simulation investigated the performance of the parametric bootstrap as a tool for diagnosing finite sample bias caused by collinearity between A and W, with the following objectives: 1) investigate further the utility of the parametric bootstrap in a setting in which estimators could not be assumed to be asymptotically normally distributed; 2) illustrate how use of a data-adaptive approach to fit Q_n can result in a poorly performing diagnostic tool unless specific measures are taken to ensure the bootstrapped data retains the sparsity present in the original data; and 3) investigate whether inclusion of the propensity score $g_n(1|W)$ as a covariate in Q_n improved the sensitivity of the diagnostic in the setting of collinearity.

Table 8: Performance of estimators by specification in Simulation 3: g_0 in [0.001,1], shown for unbounded g_n only.

	Qcgc			**Qdgd1**		
	Bias	Var	MSE	Bias	Var	MSE
G-COMP	0.133	0.038	0.055	0.212	0.027	0.072
IPTW	0.233	0.230	0.284	0.232	0.231	0.284
A-IPTW	0.134	0.038	0.055	0.175	0.027	0.057
TMLE	0.291	0.120	0.205	0.329	0.136	0.245

Table 8 demonstrates that all estimators exhibited substantial bias, even when $\bar{Q}_n$ and g_n were consistent. This remained true regardless of the level at which g_n was bounded; in the interest of space, results across bounding levels for g_n are not shown for this simulation. When stepwise selection was used to estimate Q_0, (forcing inclusion of A), the algorithm did not select $W2$ due to the collinearity with A. The consequences are reflected in the greater bias of $Qdgd1$ versus $Qcgc$ in those estimators that rely on Q_0.

Table 9: True finite sample bias for G-computation, IPTW and A-IPTW estimators (based on 250 samples of size 1000 with Qcgc) and mean and variance of estimated $ETA.Bias$ (based on the first 10 of the 250 samples) by specification in Simulation 3: g_0 in [0.001,1], shown for unbounded g_n only.

G-COMP	True finite sample bias	1.33e-01
Qcgc	Mean(ETA.Bias)	4.18e-02
	Variance(ETA.Bias)	5.62e-03
	Mean(ETA.Bias)/True Bias	3.14e-01
Stepwise G-COMP	True finite sample bias	2.12e-01
Qdgd1	Mean(ETA.Bias)	1.97e-02
	Variance(ETA.Bias)	1.21e-03
	Mean(ETA.Bias)/True Bias	9.29e-02
Qdgd2	Mean(ETA.Bias)	1.17e-01
	Variance(ETA.Bias)	1.37e-02
	Mean(ETA.Bias)/True Bias	5.52e-01
IPTW	True finite sample bias	2.33e-01
Qcgc	Mean(ETA.Bias)	8.19e-02
	Variance(ETA.Bias)	4.89e-03
	Mean(ETA.Bias)/True Bias	3.51e-01
Stepwise IPTW	True finite sample bias	2.32e-01
Qdgd1	Mean(ETA.Bias)	7.03e-02
	Variance	5.44e-03
	Mean(ETA.Bias)/True Bias	3.03e-01
Qdgd2	Estimated ETA.Bias	1.41e-01
	Variance(ETA.Bias)	1.34e-02
	Mean(ETA.Bias)/True Bias	6.08e-01
A-IPTW	True finite sample bias	1.34e-01
Qcgc	Mean(ETA.Bias)	4.20e-02
	Variance(ETA.Bias)	5.63e-03
	Mean(ETA.Bias)/True Bias	3.14e-01
Stepwise A-IPTW	True finite sample bias	1.75e-01
Qdgd1	Mean(ETA.Bias)	1.47e-02
	Variance(ETA.Bias)	7.14e-04
	Mean(ETA.Bias)/True Bias	8.40e-01
Qdgd2	Mean(ETA.Bias)	9.66e-02
	Variance	1.22e-02
	Mean(ETA.Bias)/True Bias	5.52e-01

Table 10: True finite sample bias for TMLE estimators (based on 250 samples of size 1000 with Qcgc) and mean and variance of estimated $ETA.Bias$ (based on the first 10 of the 250 samples) by specification in Simulation 3: g_0 in [0.001,1], shown for unbounded g_n only.

TMLE	True finite sample bias	2.91e-01
Qcgc	Mean(ETA.Bias)	1.70e-01
	Variance(ETA.Bias)	1.05e-02
	Mean(ETA.Bias)/True Bias	5.83e-01
Stepwise TMLE	True finite sample bias	3.29e-01
Qdgd1	Mean(ETA.Bias)	1.93e-01
	Variance(ETA.Bias)	1.24e-02
	Mean(ETA.Bias)/True Bias	5.87e-01
Qdgd2	Mean(ETA.Bias)	2.56e-01
	Variance(ETA.Bias)	1.53e-02
	Mean(ETA.Bias)/True Bias	7.78e-01

Table 11: IPTW estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 3: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$	$\hat{\Psi}_{IPTW}$	SE	$ETA.Bias$
1	0.549	0.109	0.156	0.611	0.104	0.258	0.610	0.088	0.261	0.533	0.068	0.239
2	0.589	0.052	0.083	0.594	0.052	0.160	0.534	0.052	0.144	0.421	0.050	0.123
3	0.095	0.156	0.087	0.391	0.120	0.159	0.454	0.090	0.167	0.460	0.065	0.148
4	0.740	0.050	0.090	0.751	0.050	0.067	0.761	0.050	0.026	0.670	0.051	-0.041
5	0.579	0.117	0.191	0.648	0.106	0.278	0.708	0.089	0.297	0.578	0.066	0.236
6	0.753	0.050	0.079	0.765	0.050	0.019	0.626	0.051	-0.078	0.494	0.051	-0.164
7	0.154	0.074	0.020	0.209	0.076	0.056	0.350	0.082	0.060	0.375	0.068	0.040
8	0.831	0.050	0.018	0.667	0.050	-0.058	0.526	0.050	-0.156	0.444	0.050	-0.224
9	1.147	0.048	-0.043	0.648	0.048	-0.102	0.557	0.048	-0.167	0.521	0.048	-0.230
10	0.043	0.154	0.138	0.492	0.109	0.238	0.592	0.081	0.243	0.538	0.062	0.214
Mean	0.548	0.096	0.082	0.578	0.082	0.107	0.572	0.070	0.080	0.503	0.059	0.034

Table 12: TMLE estimate, standared error and ETA.Bias estimate by sample and by bound on g_n with Qcgc, in Simulation 3: g_0 in [0.001,1]

	None			**[0.025,0.975]**			**[0.05,0.95]**			**[0.1,0.9]**		
	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$	$\hat{\Psi}_{TMLE}$	SE	$ETA.Bias$
1	0.281	0.179	0.242	0.281	0.179	0.251	1.000	0.321	0.385	1.000	0.239	0.574
2	0.263	0.179	0.250	0.263	0.179	0.198	0.263	0.179	0.224	1.000	0.242	0.463
3	0.308	0.182	0.255	0.308	0.182	0.223	0.308	0.182	0.296	1.000	0.241	0.461
4	0.330	0.184	0.003	0.680	0.164	0.021	0.683	0.164	0.046	1.000	0.176	0.177
5	0.420	0.187	0.243	0.336	0.184	0.242	1.000	0.322	0.368	1.000	0.243	0.458
6	1.000	0.167	0.194	1.000	0.167	0.196	1.000	0.170	0.259	1.000	0.176	0.350
7	0.328	0.184	0.019	0.328	0.184	0.026	0.328	0.184	0.033	0.328	0.187	0.107
8	0.286	0.179	0.251	1.000	0.170	0.248	1.000	0.173	0.299	1.000	0.179	0.341
9	1.000	0.167	0.059	0.739	0.164	0.058	1.000	0.170	0.083	1.000	0.173	0.211
10	0.379	0.192	0.180	0.319	0.190	0.138	0.319	0.190	0.194	1.000	0.247	0.342
Mean	0.459	0.179	0.170	0.525	0.176	0.160	0.690	0.232	0.219	0.933	0.217	0.348

The parametric bootstrap underestimated $ETA.Bias$ more substantially in this simulation. It would have provided a reasonable albeit imperfect diagnostic tool. Tables 9 and 10 demonstrate that for all estimators, when $\bar{Q}_n$ and g_n were consistent the estimates of $ETA.Bias$ captured 30-35% of the true finite sample bias of the G-computation, IPTW, and A-IPTW estimators , and 58% of the finite sample bias of the TMLE estimator. Tables 11 and 12 show the sample-specific $ETA.Bias$ estimates for the IPTW and TMLE estimators. When compared by an analyst to the corresponding point and variance estimates for the target parameter, the diagnostic would have suggested caution in many but not all cases. Tables 9 and 10 further demonstrate that use of a stepwise algorithm that forces A to be included in $\bar{Q}_n$ generally resulted in a greater underestimate of $ETA.Bias$ because bootstrap data are simulated from a distribution in which sparsity plays less of a role. Retention of the propensity score in the fit of $\bar{Q}_0$ that was used to generate the bootstrap data ($Qdgd2$) improved the sensitivity of the diagnostic.

5.6 Discussion of Simulation Results.

In summary, examination of the estimated treatment mechanism and corresponding propensity scores $g(a|W)$ may provide an initial alert to the presence of positivity violations; however, this approach does not provide a quantitative estimate of the resulting bias. The parametric bootstrap is a supplemental tool that allows the analyst to evaluate estimator behavior under a range of hypothetical data-generating distributions in which both the true value of the target parameter and the correct specification of nuisance parameter models is known. Further study of the performance of the diagnostic under a range of true and estimated data generating distributions is needed.

6 Data example: HIV resistance mutations.

6.1 Data and question.

We analyzed an observational cohort of HIV-infected patients in order to estimate the effect of mutations in the HIV protease enzyme on viral response to the antiretroviral drug lopinavir. The question, data, and analysis have been described previously.[33] Here, a simplified version of prior analyses was performed and the parametric bootstrap was applied to investigate the potential impact of positivity violations on results.

Briefly, baseline covariates, mutation profiles prior to treatment change, and viral response to therapy were collected for 401 treatment change episodes (TCEs) in which protease inhibitor-experienced subjects initiated a new antiretroviral regimen containing the drug lopinavir. We focused on 2 target mutations in the protease

enzyme: p82AFST and p82MLC (present in 25% and 1% of TCEs, respectively). The data for each target mutation consisted of $O = (W, A, Y)$, where A was a binary indicator that the target mutation was present prior to treatment change, W was a set of 35 baseline characteristics including summaries of past treatment history, mutations in the reverse transcriptase enzyme, and a genotypic susceptibility score for the background regimen (based on the Stanford scoring system; http://hivdb.stanford.edu/). The outcome Y was the change in $\log_{10}$(viral load) following initiation of the new antiretroviral regimen. The target observed data parameter was $E_W(E(Y|A = 1, W) - E(Y|A = 0, W))$, equal under (7) to the average treatment effect $E(Y_1 - Y_0)$.

6.2 Methods.

Effect estimates were obtained for each mutation using the IPTW estimator and TMLE with a logistic fluctuation.[34] $\bar{Q}_0$ and g_0 were estimated with stepwise forward selection of main terms based on the AIC criterion, using the `step` function in the *stats* v2.11.1 package in R. Estimators were implemented using both unbounded values for $g_n(A \mid W)$ and values truncated at $[0.025, 0.975]$. Following standard practice in much of the literature, standard errors were estimated using the influence curve, corresponding to the standard output for the `glm` and `tmle` functions in R, treating the values of g_n as fixed. The parametric bootstrap was used to estimate bias for each estimator using 1000 samples and the $ETA.Bias$ algorithm, with the `step` function rerun in each parametric bootstrap sample.

6.3 Results.

Results for both mutations are presented in Table 13. p82AFST is known to be a major mutation for lopinavir resistance.[35] The current results support this finding; the IPTW and TMLE point estimates were similar and both suggested a significantly more positive change in viral load (corresponding to a less effective drug response) among subjects with the mutation as compared to those without it. The parametric bootstrap-based bias estimate was minimal, raising no red flag that these findings might be attributable to positivity bias.

The role of mutation p82CLM is less clear based on existing knowledge; depending on the scoring system used it is either not considered a lopinavir resistance mutation, or given an intermediate lopinavir resistance score (http://hivdb.stanford.edu/).[35] Initial inspection of the point estimates and standard errors in the current analysis would have suggested that p82CLM had a large and highly significant effect on lopinavir resistance. Application of the parametric bootstrap-based diagnostic, however, would have suggested that these results should be interpreted with caution. In particular, the bias estimate for the unbounded TMLE was larger than the estimated standard error, while the bias estimate for the unbounded IPTW estimator was of roughly the

Table 13: Point estimate, standard error and parametric bootstrap-based bias estimates for the effect of two HIV resistance mutation on viral response, by estimator and bound on g_n.

	TMLE Estimator			**IPTW Estimator**		
	$\hat{\beta}_{TMLE}$	$\hat{SE}$	$ETA.Bias$	$\hat{\beta}_{IPTW}$	$\hat{SE}$	$ETA.Bias$
p82AFST						
$[0, 1]$	0.65	0.13	-0.01	0.66	0.15	-0.01
$[0.025, 0.975]$	0.62	0.13	0.00	0.66	0.15	-0.01
p82MLC						
$[0, 1]$	2.85	0.14	-0.37	1.29	0.14	0.09
$[0.025, 0.975]$	0.86	0.10	-0.01	0.80	0.23	0.08

same magnitude. While neither bias estimate was of sufficient magnitude relative to the point estimate to change inference, their size relative to the corresponding standard errors would have suggested that further investigation was warranted.

In response, the non-parametric bootstrap (based on 1000 bootstrap samples) was applied to provide an alternative estimate of the standard error. Using this alternative approach, the standard errors for the unbounded TMLE and IPTW estimators of the effect of p82MLC were estimated to be 2.77 and 1.17, respectively. Non-parametric bootstrap-based standard error estimates for the bounded TMLE and IPTW estimators were lower (0.84 and 1.12, respectively), but still substantially higher than the initial naive standard error estimates. These revised standard error estimates dramatically changed interpretation of results, suggesting that the current analysis was unable to provide essentially any information on the presence, magnitude, or direction of the p82CLM effect. (Non-parametric bootstrap-based standard error estimates for p82AFST were also somewhat larger than initial estimates, but did not change inference).

In this example, $ETA.Bias$ is expected to include some non-positivity bias due to the curse of dimensionality. However, the resulting bias estimate should still be interpreted as highly optimistic (i.e. as an underestimate of the true finite sample bias). The parametric bootstrap sampled from estimates of g_0 and $\bar{Q}_0$ that had been fit using the `step` algorithm. This ensured that the estimators g_n and $\bar{Q}_n$ (which applied the same stepwise algorithm) would do a good job approximating $g_{\hat{P}_0}$ and $\bar{Q}_{\hat{P}_0}$ in each bootstrap sample. Clearly, no such guarantee exists for the true P_0. This simple example further illustrates the utility of the non-parametric bootstrap for standard error estimation in the setting of sparse data and positivity violations. In this particular example, the improved variance estimate provided by the non-parametric bootstrap was sufficient to prevent positivity violations from leading to incorrect inference. As demonstrated in the simulations, however, in other settings even accurate variance estimates may fail to alert the analyst to threats posed by positivity violations.

7 Practical approaches to causal inference in the presence of positivity violations

How should analysis proceed once threats to inference due to data sparsity have been identified? In this section we review several approaches to effect estimation in the presence of positivity violations. These include changing the projection function $h(a, V)$ used to defined the target parameter β, restricting the covariate adjustment set, restricting the sample, and redefining the causal effect of interest through the use of realistic and intention to treat parameters. Moore *et. al.* provide an extended review of these approaches.[15] All four approaches can be viewed as a means to define a family of parameters that approximate the original target of inference to differing degrees. Estimators can then be defined that select among members of a given family based on the tradeoff between degree of divergence from the original target and identifiability.

7.1 Approach #1: Change the projection function $h(A, V)$.

Throughout this paper we have focused on the target causal parameter $\beta(F_X, m, h)$ defined according to (5) as the projection of the $E_{F_X}(Y_a|V)$ on the marginal structural model $m(a, V|\beta)$. Choice of function $h(a, V)$ both defines the target parameter by specifying which values of (A, V) should be given greater weight when estimating β and, by assumption (9), defines the positivity assumption needed for β to be identifiable.

We have focused on parameters indexed by $h(a, V) = 1$, a choice that gives equal weight to estimating the counterfactual outcome for all values (a, v).[11] Alternative choices of $h(a, V)$ can significantly weaken the needed positivity assumption. For example, if the target of inference only involves counterfactual outcomes among some restricted range $[c, d]$ of possible values $\mathcal{A}$, defining $h(a, V) = I(a \in [c, d])$ weakens the positivity assumption by requiring sufficient variability only in the assignment of treatment levels within the target range. In some settings, the causal parameter defined by such a projection over a limited range of $\mathcal{A}$ might be of substantial *a priori* interest. For example, one may wish to focus estimation of a drug dose response curve only on the range of doses considered reasonable for routine clinical use, rather than on the full range of doses theoretically possible or observed in a given data set.

An alternative approach, commonly employed in the context of IPTW estimation and introduced in Section 3.2, is to choose $h(a, V) = g(a|V)$, where $g(a|V) \equiv P(A = a|V)$ is the conditional probability of treatment given the covariates included in the marginal structural model. In the setting of IPTW estimation this choice corresponds to the use of stabilizing weights, a common approach to reducing both the variance of the IPTW estimator in the face of sparsity.[21] When the target causal parameter is defined using a non-parametric marginal structural model, use of $h(a, V) = g(a, V)$ corresponds with a decision to define a target parameter that gives greater weight

to those regions of the joint distribution of (A, V) that are well-supported, and that relies on smoothing or extrapolation to a greater degree in areas that are not.[11]

Use of a marginal structural working model makes clear that the utility of choosing $h(a, V) = g(a|V)$ as a method to approach data sparsity is not limited to the IPTW estimator. Recall that the G-computation estimator can be implemented by regressing predicted values for Y_a on (a, V) according to model $m(a, V|\beta)$ with weights provided by $h(a, V)$. When the projection function is chosen to be $g(a|V)$, this corresponds to a weighted regression in which weights are proportional to the degree of support in the data.

Even when one is ideally interested in the entire causal curve (implying a target parameter defined by choice $h(a, V) = 1$), specification of alternative choices for h offers a means of improving identifiability, at a cost of redefining the target parameter. For example, one can define a family of target parameters indexed by $h_\delta(a, V) = I(a \in [c(\delta), d(\delta)])$, where an increase in δ corresponds to progressive restriction on the range of treatment levels targeted by estimation. Fluctuation of δ thus corresponds to trading a focus on more limited areas of the causal curve for improved parameter identifiability. Selection of the final target from among this family can be based on an estimate of bias provided by the parametric bootstrap. For example, the bootstrap can be used to select the parameter with the smallest δ below some pre-specified threshold for allowable $ETA.Bias$.

7.2 Approach #2: Restrict the adjustment set.

Exclusion of problematic W (i.e. those covariates resulting in positivity violations or near violations) from the adjustment set, provides a means to trade confounding bias for a reduction in positivity violations.[36] In some cases, exclusion of covariates from the adjustment set may come at little or no cost to bias in the estimate of the target parameter. In particular, a subset of W that excludes covariates responsible for positivity violations may still be sufficient to control for confounding. In other words, a subset $W' \subset W$ may exist for which both identifying assumptions (7) and (8) hold (i.e. $Y_a \coprod A \mid W'$ and $g_0(a|W') > 0, a \in \mathcal{A}$), while positivity fails for the full set of covariates. In practice, this approach can be implemented by first determining candidate subsets of W under which the positivity assumption holds, and then using causal graphs to assess whether any of these candidates is sufficient to control for confounding. Even when no such candidate set can be identified, background knowledge (or sensitivity analysis) may suggest that problematic W represent a minimal source of confounding bias (Moore et. al. provide an example).[15] Often, however, those covariates that are most problematic from a positivity perspective are also strong confounders.

As suggested with respect to choice of projection function $h(a, V)$ in the previous section, the causal effect estimator can be fine-tuned to select the degree of restriction on the adjustment set W according to some pre-specified rule for eliminating covariates

from the adjustment set, and the parametric bootstrap used to select the minimal degree of restriction that maintains $ETA.Bias$ below an acceptable threshold.[36] Also, the C-TMLE estimator mentioned briefly in Section 3.3, which includes in the fit of g_n only those covariates that improve estimation of the target parameter, will restrict W in a "black-box" manner. In the case of substantial positivity violations, such approaches can result in small covariate adjustment sets. While such limited covariate adjustment accurately reflects a target parameter that is poorly supported by the available data, the resulting estimate can be difficult to interpret and will no longer carry a causal interpretation.

7.3 Approach # 3: Restrict the sample.

An alternative approach, sometimes referred to as "trimming", is to discard classes of subjects for whom there exists no or limited variability in observed treatment assignment. A causal effect is then estimated in the remaining subsample. This approach is popular in the econometrics and social science literature; Crump provides a recent review.[37;38;39;40]

When the subset of covariates responsible for positivity violations is low or one dimensional, such an approach can be implemented simply by discarding subjects with covariate values not represented in all treatment groups. For example, say that one aims to estimate the average effect of a binary treatment, and in order to control for confounding needs to adjust for W, a covariate with possible levels $\{1, 2, 3, 4\}$. However, inspection of the data reveals that no one in the sample with $W = 4$ received treatment (ie. $g_n(1|W = 4) = 0$). The sample can be trimmed by excluding those subjects for whom $W = 4$ prior to applying a given causal effect estimator for the average treatment effect. As a result, the target parameter is shifted from $E(Y_1 - Y_0)$ to $E(Y_1 - Y_0|W < 4)$, and the positivity assumption (8) now holds (as $W = 4$ occurs with zero probability).

Often W is too high dimensional to make this straightforward implementation feasible; in such a case matching on the propensity score provides a means to trim the sample. There is an extensive literature on propensity score-based effect estimators; however such estimators are beyond the scope of the current review. Several potential problems arise with the use of trimming methods to address positivity violations. First, discarding subjects responsible for positivity violations shrinks sample size, and thus runs the risk of increasing the variance of the effect estimate. Further, sample size and the extent to which positivity violations arise by chance are closely related. Depending on how trimming is implemented, new positivity violations can be introduced as sample size shrinks. Second, restriction of the sample may result in a causal effect for a population of limited interest. In other words, as can occur with alternative approaches to improve identifiability by shifting the target of inference, the parameter actually estimated may be far from the initial target. Further, when the criterion used to restrict the sample involves a summary of high dimensional covariates, such as is provided the propensity score, it can be difficult to interpret

the parameter estimated. Finally when treatment is longitudinal, the covariates responsible for positivity violations may themselves be affected by past treatment.[15] Trimming to remove positivity violations in this setting amounts to conditioning on post-treatment covariates and can thus introduce new bias.

Crump proposes an approach to trimming that falls within the general strategy of redefining the target parameter in order to explicitly capture the tradeoff between parameter identifiability and proximity to the initial target.[37] In addition to focusing on the treatment effect in an *a priori* specified target population, he defines an alternative target parameter corresponding to the average treatment effect in that subsample of the population for which the most precise estimate can be achieved. Crump further suggests the potential for extending this approach to achieve an optimal (according to some user-specified criteria) tradeoff between the representativeness of the subsample in which the effect is estimated and the variance of the estimate.

7.4 Approach #4: Change the intervention of interest.

A final alternative for improving the identifiability of a causal parameter in the presence of positivity violations is to redefine the intervention of interest. Realistic rules rely on an estimate of the propensity score $g(a|W)$ to define interventions that explicitly avoid positivity violations. This ensures that the causal parameter estimated is sufficiently supported by existing data.

Realistic interventions avoid positivity violations by first identifying subjects for whom a given treatment assignment is not realistic (i.e. subjects whose propensity score for a given treatment is small or zero) and then assigning an alternative treatment with better data support to those individuals. Such an approach is made possible by focusing on the causal effects of dynamic treatment regimes.[41;42] The causal parameters described thus far are summaries of the counterfactual outcome distribution under a fixed treatment applied uniformly across the target population. In contrast, a dynamic regime assigns treatment in response to patient covariate values. This characteristic makes it possible to define interventions under which a subject is only assigned treatments that are possible (or "realistic") given a subject's covariate values.

To continue the previous example in which no subjects with $W = 4$ were treated, a realistic treatment rule might take the form "treat only those subjects with W less than 4." More formally, let $d(W)$ refer to a treatment rule that deterministically assigns a treatment $a \in \mathcal{A}$ based on a subject's covariates W and consider the rule $d(W) = I(W < 4)$. Let Y_d denote the counterfactual outcome under the treatment rule $d(W)$, which corresponds to treating a subject if and only if his or her covariate W is below 4. In this example $E(Y_0)$ is identified as $\sum_w E(Y|W = w, A = 0)P(W = w)$; however, since $E(Y|W = w, A = 1)$ is undefined for $W = 4$, $E(Y_1)$ is not identified (unless we are willing to extrapolate based on $W < 4$). In contrast, $E(Y_d)$ is identified by the non-parametric G-computation formula: $\sum_w E(Y = y|W = w, A =$

$d(W))P(W = w)$. Thus the average treatment effect $E(Y_d - Y_0)$, but not $E(Y_1 - Y_0)$, is identified. The redefined causal parameter can be interpreted as the difference in expected counterfactual outcome if only those subjects with $W < 4$ were treated as compared to the outcome if no one were treated.

More generally, realistic rules indexed by a given static treatment a assign a only to those individuals for whom the probability of receiving a is greater than some user-specified probability α (such as $\alpha > 0.05$). Let $d(a, W)$ denote the rule indexed by static treatment a. If A is binary, then $d(1, W) = 1$ if $g(1|W) > \alpha$, otherwise $d(1, W) = 0$. Similarly, $d(0, W) = 0$ if $g(0|W) > \alpha$; otherwise $d(0, W) = 1$. Realistic causal parameters are defined as some parameter of the distribution of $Y_{d(a,W)}$ (possibly conditional on some subset of baseline covariates $V \subset W$). Estimation of the causal effects of dynamic rules $d(W)$ allows the positivity assumption to be relaxed to $g(d(W)|W) > 0$ -a.e (i.e. only those treatments that would be assigned based on rule d to patients with covariates W need to occur with positive probability within strata of W). Realistic rules $d(a, W)$ are designed to satisfy this assumption by definition.

When a given treatment level a is unrealistic (i.e. when $g(a \mid W) < \alpha$), realistic rules assign an alternative from among viable (well-supported) choices. Choice of an alternative is straightforward when treatment is binary. When treatment has more than two levels, however, a rule for selecting the alternative treatment level is needed. One option is to assign a treatment level that is as close as possible to the orignal assignment while still remaining realistic. For example, if high doses of drugs occur with low probability in a certain subset of the population, a realistic rule might assign the maximum dose that occurs with probability $> \alpha$ in that subset. An alternative class of dynamic regimes, referred to as "intent-to-treat" rules, instead assign a subject to his or her observed treatment value if an initial assignment is deemed unrealistic. Moore, *et. al.* and Bembom, *et. al.* provide illustrations of both of these types of realistic rules using simulated and real data.[15;14]

The causal effects of realistic rules clearly differ from their static counterparts. The extent to which the new target parameter diverges from the initial parameter of interest depends on both the extent to which positivity violations occur in the finite sample (i.e. the extent of support available in the data for the initial target parameter) and on a user-supplied threshold α. The parametric bootstrap approach presented in Section 4 can be employed to data-adaptively select α based on the level of $ETA.Bias$ deemed acceptable.[14]

7.5 Selection among a family of parameters.

Each of the methods described for estimating causal effects in the presence of data sparsity corresponds to a particular strategy for altering the target parameter in exchange for improved identifiability. In each case, we have outlined how this trade-off could be made systematically, based on some user-specified criterion such as the

bias estimate provided by the parametric bootstrap. We now summarize this general approach in terms of a formal method for estimation in the face of positivity violations.

1. Define a family of parameters. The family should include the initial target of inference together with a set of related parameters, indexed by γ in index set I, where γ represents the extent to which a given family member trades improved identifiability for decreased proximity to the initial target. In the examples given in the previous section, γ could be used to index a set of projection functions $h(a, V)$ based on an increasingly restrictive range of the possible values $\mathcal{A}$, degree to which the adjustment covariate set or sample is restricted, or choice of a threshold for defining a realistic rule.

2. Apply the parametric bootstrap to generate an estimate $ETA.Bias$ for each $\gamma \in I$. In particular, this involves estimating the data generating distribution, simulating new data from this estimate, and then applying an estimator to each target indexed by γ.

3. Select the target parameter from among the set that fall below a pre-specified threshold for acceptable $ETA.Bias$. In particular, select the parameter from within this set that is indexed by the value γ that corresponds to the greatest proximity to the initial target.

This approach allows an estimator to be defined in terms of an algorithm that identifies and estimates the parameter within a candidate family that is as close to the initial target of inference as possible while remaining within some user-supplied limit on the extent of tolerable positivity violations.

8 Conclusions.

The identifiability of causal effects relies on sufficient variation in treatment assignment within covariate strata. The strong version of positivity requires that each possible treatment occur with positive probability in each covariate strata; depending on the model and target parameter, this assumption can be relaxed to some extent. In addition to assessing identifiability based on measurement of and control for sufficient confounders, data analyses should directly assess threats to identifiability based on positivity violations. The parametric bootstrap is a practical tool for assessing such threats, and provides a quantitative estimator-specific estimate of bias arising due to positivity violations.

The objective of the parametric bootstrap diagnostic is to raise a red flag in settings where positivity violations (as well as bounding of g_n) may be resulting in bias of sufficient magnitude to threaten reliable inference. The simulations showed that the diagnostic worked best when (1) Q_n and g_n were consistently estimated; (2) g_n was at least minimally bounded so that the estimator was more likely to be asymptotically

normal; and, (3) any data-adaptive algorithm used to fit $\bar{Q}_0$ was forced to include not only A but also the propensity score in order to retain sparsity in the bootstrapped distribution. Although the diagnostic may underestimate the true $ETA.Bias$, in the simulations presented here the diagnostic was generally successful in raising a red-flag for bias due to positivity violations in the settings where such a warning was needed. The performance of the diagnostic should be further investigated under a range of true and estimated data generating distributions, however.

This paper has focused on the positivity assumption for the causal effect of a treatment assigned at a single time point. Extension to a longitudinal setting in which the goal is to estimate the effect of multiple treatments assigned sequentially over time introduces considerable additional complexity. First, practical violations of the positivity assumption can arise more readily in this setting. Under the longitudinal version of the positivity assumption the conditional probability of each possible treatment history should remain positive regardless of covariate history. However, this probability is the product of time point-specific treatment probabilities given the past. When the product is taken over multiple time points it is easy for treatment histories with very small conditional probabilities to arise. Second, longitudinal data make it harder to diagnose the bias arising due to positivity violations. Implementation of the parametric bootstrap in longitudinal settings requires Monte Carlo simulation both to implement the G-computation estimator and to generate each bootstrap sample. In particular, this requires estimating and sampling from the time-point specific conditional distributions of all covariates and treatment given the past. Additional research on assessing the impact of of positivity bias on longitudinal causal parameters is needed, including investigation of the parametric bootstrap in this setting.

When positivity violations occur for structural reasons rather than due to chance, a causal parameter that avoids these positivity violations will often be of substantial interest. For example, when certain treatment levels are contraindicated for certain types of individuals, the average treatment effect in the population may be of less interest than the effect of treatment among that subset of the population without contraindications, or alternatively, the effect of an intervention that assigns treatment only to those subjects without contraindications. Similarly, the effect of a multilevel treatment may be of greatest interest for only a subset of treatment levels.

In other cases researchers may be happy to settle for a better estimate of a less interesting parameter. Sample restriction, estimation of realistic parameters, and change in projection function $h(a, V)$ all change the causal effect being estimated; in contrast, restriction of the covariate adjustment set often results in estimation of a non-causal parameter. However, all of these approaches can be understood as means to shift from a poorly identified initial target towards a parameter that is less ambitious but more fully supported by the available data. The new estimand is not determined *a priori* by the question of interest, but rather is driven by the observed data distribution in the finite sample at hand. There is thus an explicit tradeoff between identifiability and proximity to the initial target of inference. Ideally, this tradeoff will be made in a systematic way rather than on an *ad hoc* basis at the

discretion of the investigator. Definition of an estimator that selects among a family of parameters according to some pre-specified criteria is a means to formalize this tradeoff. An estimate of bias based on the parametric bootstrap can be used to implement the tradeoff in practice.

The parametric bootstrap also provides a means to optimize estimator performance without changing the target parameter. The parametric bootstrap provides an estimate of the whole sampling distribution of a candidate estimator, and thus can be used to estimate MSE and fine-tune estimator performance based on this estimate. Bembom *et. al.* illustrate this approach by using the bootstrap to data-adpatively select the level of weight truncation that minimizes the estimated MSE of the IPTW estimator; the same method can also be used to minimize estimated MSE using alternative approaches such as progressive restriction of the adjustment set. We emphasize, however, that use of the parametric bootstrap to minimize estimator MSE is fundamentally different than use of the parametric bootstrap to select among a family of parameters, as described in Section 7.5. The former represents a means of improving estimator performance for the same target parameter (by fine-tuning the estimator to optimize bias-variance tradeoff). In contrast, the family of parameters approach shifts the target of inference to a parameter that is adequately supported by the data.

In summary, we offer the following advice for applied analyses: First, define the causal effect of interest based on careful consideration of structural positivity violations. Second, consider estimator behavior in the context of positivity violations when selecting an estimator. Third, apply the parametric bootstrap to quantify the extent of estimator bias under data simulated to approximate the true data generating distribution. Fourth, when positivity violations are a concern, choose an estimator that selects systematically among a family of parameters based on the tradeoff between data support and proximity to the initial target of inference.

References

[1] W.G. Cochran. Analysis of covariance: Its nature and uses. *Biometrics*, 13:261–281, 1957.

[2] J.M. Robins. A new approach to causal inference in mortality studies with sustained exposure periods - application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7:1393–1512, 1986.

[3] J.M. Robins. Addendum to: "A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect" [Math. Modelling **7** (1986), no. 9-12, 1393–1512; MR 87m:92078]. *Comput. Math. Appl.*, 14(9-12):923–945, 1987.

[4] J.M. Robins. Robust estimation in sequentially ignorable missing data and causal

inference models. In *Proceedings of the American Statistical Association: Section on Bayesian Statistical Science*, pages 6–10, 1999.

[5] Y. Wang, M. Petersen, D. Bangsberg, and M.J. van der Laan. Diagnosing bias in the inverse probability of treatment weighted estimator resulting from violation of experimental treatment assignment. Technical Report 211, Division of Biostatistics, University of California, Berkeley, 2006.

[6] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[7] J. Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5:465–480, 1923.

[8] D.B. Rubin. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

[9] J.M. Robins. Marginal structural models. In *Proceedings of the American Statistical Association. Section on Bayesian Statistical Science 1997*, pages 1–10, 1998.

[10] J.M. Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials (Minneapolis, MN, 1997)*, pages 95–133. Springer, New York, 1999.

[11] R. Neugebauer and M. J. van der Laan. Non-parametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137(2):419–434, 2007.

[12] J.M. Robins. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *Journal of Chronic Disease*, 40(2):139s–161s, 1987.

[13] R. Neugebauer and M.J. van der Laan. Why prefer double robust estimates. *Journal of Statistical Planning and Inference*, 129(1-2):405–426, 2005.

[14] O. Bembom and M.J. van der Laan. A practical illustration of the importance of realistic individualized treatment rules in causal inference. *Electronic Journal of Statistics*, 1:574–596, 2007.

[15] K.L. Moore, R.S. Neugebauer, M.J. van der Laan, and I.B. Tager. Causal inference in epidemiological studies with strong confounding. Technical Report 255, Division of Biostatistics, University of California, Berkeley, 2009.

[16] S.R. Cole and M.A. Hernan. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology*, 168:656–664, 2008.

[17] M. M. Rosenblum and M. van der Laan. Confidence intervals for the population mean tailored to small sample sizes, with applications to survey sampling. *The International Journal of Biostatistics*, 1:4, 2001.

[18] M.J. van der Laan and S. Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. Technical Report 130, Division of Biostatistics, University of California, Berkeley, 2003.

[19] M.J. van der Laan, E. C. Polley, and A. E. Hubbard. Super learner. *Genetics and Molecular Biology*, 6, 2007.

[20] T. Hastie, R. Tibshirani, and J Friedman. *The Elements of Statistical Learning*. Springer, London, 2009.

[21] J.M. Robins, M.A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.

[22] L. Kish. Weighting for unequal p_i. *Journal of Official Statistics*, 8:183–200, 1992.

[23] O. Bembom and M.J. van der Laan. Data-adaptive selection of the truncation level for inverse-probability-of-treatment-weighted estimators. Technical Report 230, Division of Biostatstics, University of California, Berkeley, 2008.

[24] J. M. Robins and A. Rotnitzky. Comment on the Bickel and Kwon article, "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica*, 11(4):920–936, 2001.

[25] J. M. Robins. Commentary on using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard by Dawson and Lavori". *Statistics in Medicine*, 21:1663–1680, 2002.

[26] D.O. Scharfstein, A. Rotnitzky, and J.M. Robins. Adjusting for non-ignorable drop-out using semiparametric nonresponse models, (with discussion and rejoinder). *Journal of the American Statistical Association*, 94:1096–1120 (1121–1146), 1999.

[27] M.J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1):11, 2006.

[28] M. Rosenblum and M. van der Laan. Targeted maximum likelihood estimation of the parameter of a marginal structural model. *The International Journal of Biostatistics*, 6(2), 2010.

[29] M.J. van der Laan and S. Gruber. Collaborative double robust targeted penalized maximum likelihood estimation. Technical Report 246, Division of Biostatistics, University of California, Berkeley, 2009.

[30] S. Gruber and M.J. van der Laan. Estimator of a causal effect on a bounded continuous outcome. Technical Report 265, U.C. Berkeley Division of Biostatistics Working Paper Series., 2000.

[31] D.A. Freedman and R.A. Berk. Weighting regressions by propensity scores. *Evaluation Review*, 32(4):392–409, 2008.

[32] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Emprical Processes*. Springer-Verlag New York, 1996.

[33] O. Bembom, M.L. Petersen, S.-Y. Rhee, W. J. Fessel, S.E. Sinisi, R.W. Shafer, and M.J. van der Laan. Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant HIV infection. *Statistics in Medicine*, 28:152–72, 2009.

[34] S. Gruber and M.J. van der Laan. An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6(1):Article 18, 2010.

[35] V.A. Johnson, F. Brun-Vezinet, and et. al. B. Clotet. Update of the drug resistance mutations in HIV-1: December 2009. *Topics in HIV Medicine*, 17(5):138–45, 2009.

[36] O. Bembom, J.W. Fessel, R.W. Shafer, and M.J. van der Laan. Data-adaptive selection of the adjustment set in variable importance estimation. Technical Report 231, Division of Biostatstics, University of California, Berkeley, 2008.

[37] R.K. Crump, V.J. Hotz, G.W. Imbens, and O.A. Mitnik. Moving the goalposts: Adressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical Report 330, National Bureau of Economic Research, 2006.

[38] R.J. LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review*, 76:604–620, 1986.

[39] J. Heckman, H. Ichimura, and R. Todd. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Review of Economic Studies*, 64:605–654, 1997.

[40] R. Dehejia and S.Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94:1053–1062, 1999.

[41] M.J. van der Laan and M.L. Petersen. Causal effect models for realistic individualized treatment and intention to treat rules. *The International Journal of Biostatistics*, 3(1):3, 2007.

[42] J.M. Robins, L. Orellana, and Andrea Rotnitzky. Estimation and extrapolation of optimal treatment and testing strategies. *Statistics in Medicine*, 27:4678–4721, 2008.