

Didelez, Vanessa

**Article — Published Version**

Seconder of the vote of thanks to Vansteelandt and Dukes and contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters'

Journal of the Royal Statistical Society: Series B (Statistical Methodology)

**Provided in Cooperation with:**

John Wiley & Sons

*Suggested Citation:* Didelez, Vanessa (2022) : Seconder of the vote of thanks to Vansteelandt and Dukes and contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters', Journal of the Royal Statistical Society: Series B (Statistical Methodology), ISSN 1467-9868, Wiley, Hoboken, NJ, Vol. 84, Iss. 3, pp. 689-691, <https://doi.org/10.1111/rssb.12514>

This Version is available at:

<https://hdl.handle.net/10419/265003>

**Standard-Nutzungsbedingungen:**

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

**Terms of use:**

*Documents in EconStor may be saved and copied for your personal and scholarly purposes.*

*You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.*

*If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.*



<http://creativecommons.org/licenses/by-nc-nd/4.0/>

## ORIGINAL ARTICLE

## Discussion Paper

# Assumption-lean inference for generalised linear model parameters

Stijn Vansteelandt<sup>1,2</sup>  | Oliver Dukes<sup>1</sup> <sup>1</sup>Ghent University, Ghent, Belgium<sup>2</sup>London School of Hygiene and Tropical Medicine, London, UK**Correspondence**Stijn Vansteelandt, Ghent University,  
Ghent, Belgium.Email: [stijn.vansteelandt@ugent.be](mailto:stijn.vansteelandt@ugent.be)**Abstract**

Inference for the parameters indexing generalised linear models is routinely based on the assumption that the model is correct and a priori specified. This is unsatisfactory because the chosen model is usually the result of a data-adaptive model selection process, which may induce excess uncertainty that is not usually acknowledged. Moreover, the assumptions encoded in the chosen model rarely represent some a priori known, ground truth, making standard inferences prone to bias, but also failing to give a pure reflection of the information that is contained in the data. Inspired by developments on assumption-free inference for so-called projection parameters, we here propose novel nonparametric definitions of main effect estimands and effect modification estimands. These reduce to standard main effect and effect modification parameters in generalised linear models when these models are correctly specified, but have the advantage that they continue to capture respectively the (conditional) association between two variables, or the degree to which two variables interact in their association with outcome, even when these models are misspecified. We achieve an assumption-lean

[Read before The Royal Statistical Society at the Discussion Meeting held online on Tuesday, 6 July 2021, Professor Guy Nason in the Chair]

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

inference for these estimands on the basis of their efficient influence function under the nonparametric model while invoking flexible data-adaptive (e.g. machine learning) procedures.

#### KEYWORDS

bias, conditional treatment effect, estimand, influence function, interaction, model misspecification, nonparametric inference

## 1 | INTRODUCTION

Statistical analyses routinely invoke modelling assumptions. These include smoothness assumptions, implied by parametric or semiparametric model specifications, for instance, but also sparsity assumptions that underlie variable selection procedures. Such assumptions are generally a necessity. The curse of dimensionality indeed forces one to borrow information across strata of subjects with different covariate values, as well as to reduce the dimensions of the possibly many measured variables. Modelling assumptions are often also a deliberate choice. With a continuous exposure, for instance, one would often not be interested in knowing exactly how the outcome changes with each increase in exposure, but might content oneself with a ‘simple’ and parsimonious summary of the exposure effect. Models enable one to create such summaries. This distinction in the nature of the assumptions is rarely made in how we approach a data analysis, but is nonetheless an essential one that will turn out to be key to the strategy that we advocate.

Regardless of this distinction, modelling assumptions are almost always a pure mathematical convenience, and not reflecting a priori knowledge that we had prior to seeing the data. Ideally, in such cases, data analyses should therefore only extract information from the data, and not from the assumptions. This realisation is not new. It became very dominant in the 90s in work on non-ignorable incomplete data. Rotnitzky and Robins (e.g. Rotnitzky & Robins, 1997; Rotnitzky et al., 1998; Scharfstein et al., 1999), amongst others, then increased awareness that modelling assumptions, such as normality and linearity assumptions, may sometimes permit identification of parameters in the absence of missing data assumptions. There is now a fairly general agreement that such identification is dishonest when these modelling assumptions are made for convenience. In spite of this, once we have stated structural assumptions (e.g. missing data assumptions) needed for identification, we often fall back into our routine. We continue to rely on modelling assumptions more than we may realise, and treat them as representing some ground truth in how we approach inference.

For instance, likelihood-based or semiparametric estimation approaches extract information not only from the data, but also from the model as if it were known to contain the truth. In fact, maximum likelihood estimators, maximum a posteriori estimators and semiparametric efficient estimators precisely succeed to increase efficiency by taking modelling assumptions as given, and extracting information from them. This makes the resulting data analysis no longer purely evidence based. We usually try to make up for this by adopting model or variable selection procedures. However, the inferences that are commonly provided, continue to pretend that the model delivered by these procedures was a priori given and known, which can sometimes make things worse. All of this is raising questions to what extent the data analyses that we produce are effectively (purely) evidence based.

Motivated by these concerns, enormous progress has been made over the past several decades in terms of how to develop an inference that is ‘assumption-free’, across several different literatures. White (1980) developed the so-called ‘sandwich estimator’ of the standard error for ordinary least squares (OLS); this delivers a valid measure of uncertainty around the regression coefficient estimates, even if key model-based assumptions of OLS (linearity, homoscedasticity) are not met. Freedman (2006) noted that although the sandwich estimator is unbiased under nonlinearity, the resulting confidence intervals and tests are not useful given that it may be unclear what the model coefficients represent. Several proposals for restoring meaning to regression estimates have been made, seeing a model coefficient as a projection parameter (Buja et al., 2019a,b,c; Kennedy et al., 2019; van der Laan & Rose, 2011; Neugebauer & van der Laan, 2007), or variable importance measure (Chambaz et al., 2012), both ideas which have gained traction in high-dimensional statistics (Berk et al., 2013; Wasserman, 2014). In terms of doing causal inference, Lin (2013) gave a ‘model-agnostic’ approach to the adjustment for baseline covariates in randomised experiments. He noted that ‘one does not need to believe in the classical linear model to tolerate or even advocate OLS adjustment’. Related work has explored how OLS estimates can in certain settings be interpreted as weighted averages of treatment effects, even when the linear model is wrong (Angrist & Krueger, 1999; Angrist & Pischke, 2009; Aronow & Samii, 2016; Graham & Pinto, 2018; Słoczyński, 2020). Many of the above approaches start with a common estimator of a parameter indexing a parametric regression model. They then characterise to what estimand (i.e. functional of the data distribution) the estimator converges, without assuming that the model is true. In contrast, Mark van der Laan and collaborators take an alternative approach in their scientific ‘roadmap’ (van der Laan & Rose, 2011; van der Laan & Rubin, 2006). They first define an estimand which characterises what we aim to infer from the data, and next develop estimation and inference based on its efficient influence function (provided the estimand is pathwise-differentiable under the nonparametric model; see Section 5), with all nuisance functionals estimated nonparametrically (e.g. via machine learning). Reliance on the efficient influence function is essential to this development, as it enables valid inference even when the analysis is based on data-adaptive procedures, such as machine learning, variable selection, model selection, etc. Attention in their work is mainly given to causal inference applications where the focus is on the average (total or (in)direct) effect of a binary, possibly time-varying treatment on a binary or continuous outcome.

Key to the latter developments is changing the starting point of the analysis from the postulation of a statistical model to the postulation of an estimand. This change of focus brings many advantages. It forces one to work with well-understood estimands that target the scientific question from the start. It enables one to separate modelling assumptions made for parsimony, which will be used to define the estimand, from assumptions imposed to handle the curse of dimensionality. It prevents reliance on these assumptions, as inference for the estimand can be developed under the nonparametric model. Finally, the resulting analysis can be pre-specified, which is essential if one aims for an honest data analysis that reflects all uncertainties, including the uncertainty surrounding the model that is used.

Changing this focus of the analysis is non-trivial, however. It turns the difficulty of postulating a model, to which we have grown to become familiar, into the difficulty of choosing an estimand, for which infinitely many choices can typically be conceived. While there is some experience in choosing meaningful estimands in causal inference applications, complications easily arise when, for example, considering continuous exposures, or when general association measures (e.g. measures of a time trend) rather than causal effect measures are of interest. It calls for the development of specific estimands that can be used quite generically (in a sense that

we will make specific later) and connect to regression parameters that practitioners have grown to become familiar with. In this way, they can provide an assumption-lean inference for those standard regression parameters, which uses the underlying model only with the aim to summarise and to deliver a familiar interpretation, but relates to flexible statistical or machine learning procedures running in the background to assure valid inference. In this paper, we will show how we believe this is best done when the aim is to infer regression parameters indexing generalised linear models. In particular, we propose novel estimands for conditional association measures between two variables, and for the degree to which two variables interact in their association with outcome, which are well defined in a nonparametric sense (i.e. regardless of what is the underlying data-generating distribution). We achieve an assumption-lean inference for these estimands by deriving their efficient influence function under the nonparametric model and invoking flexible data-adaptive (e.g. parametric model selection or machine learning) procedures. Since the proposed estimands reduce to standard main effect and interaction parameters in generalised linear models when these models are correctly specified, we thus generalise standard inference for such parameters to give a pure reflection of the information that is contained in the data. Our developments thus provide a novel framework for fitting generalised linear models, and at a broader level, also shed light on what defines an adequate estimand, and how it can be constructed.

In Section 2, we illustrate the above concerns about parametric and semiparametric methods with a simple example. This is followed by proposals for novel main effect and interaction estimands in Sections 3 and 4, respectively. Nonparametric inference is developed for these estimands in Section 5, and the empirical performance of the resulting estimators is assessed in Section 6 via simulation studies. In Section 7 we apply our framework in an analysis of the effect of the First Steps program on infant birth-weight, before closing the paper with a discussion in Section 8.

## 2 | ILLUSTRATION

To clarify the points made in the introduction, we provide a simple illustration with artificial, independent data for  $n = 50$  subjects on a scalar standard normal variate  $L$ , a dichotomous exposure  $A$ , coded 0 or 1, with  $P(A = 1|L) = \text{expit}(L - L^2)$  and a normally distributed outcome with mean  $A - L + 4.5AL + 0.5L^2 - 2.25AL^2$  and unit (residual) variance. The ordinary least squares estimator for  $\beta$  under model

$$E(Y|A, L) = \alpha_0 + \alpha_1 L + \beta A,$$

can be shown to converge to

$$\begin{aligned} & \frac{E[\pi(L)\{1 - \tilde{\pi}(L)\}\{E(Y|A = 1, L) - E(Y|A = 0, L)\}]}{E[\pi(L)\{1 - \tilde{\pi}(L)\}]} \\ & + \frac{E[\{\pi(L) - \tilde{\pi}(L)\}E(Y|A = 0, L)]}{E[\pi(L)\{1 - \tilde{\pi}(L)\}]}, \end{aligned}$$

where  $\pi(L) = P(A = 1|L)$  is the so-called propensity score and  $\tilde{\pi}(L)$  denotes the population least squares projection of  $A$  onto 1 and  $L$ . This displayed ‘estimand’ consists of two contributions. The first is a weighted average of the contrasts  $E(Y|A = 1, L) - E(Y|A = 0, L)$ . It is informative about the conditional association between  $A$  and  $Y$ . The second contribution is a weighted average of the contrasts  $\pi(L) - \tilde{\pi}(L)$ . It is not informative about the conditional association between  $A$

and  $Y$  and is generally non-zero, except when the linear outcome model is correctly specified or  $\pi(L)$  happens to be a linear function of  $L$  (see e.g. Robins et al., 1992; Vansteelandt and Joffe, 2014). This is disturbing. It makes the estimand targeted by the ordinary least squares estimator a questionable summary of the conditional association between  $A$  and  $Y$ , given  $L$ , when the linear model is misspecified.

A more attractive approach is based on the partially linear model

$$E(Y|A, L) = \omega(L) + \beta A, \quad (1)$$

where  $\beta$  and  $\omega(L)$  are unknown. Here,  $\hat{\beta}$  can be obtained as the E-estimator

$$\frac{\sum_{i=1}^n \{A_i - \hat{\pi}(L_i)\} \{Y_i - \hat{\omega}(L_i)\}}{\sum_{i=1}^n \{A_i - \hat{\pi}(L_i)\} A_i}, \quad (2)$$

(Robins et al., 1992), where  $\hat{\pi}(\cdot)$  and  $\hat{\omega}(\cdot)$  are possibly data-adaptive estimators of  $\pi(\cdot)$  and  $\omega(\cdot)$ , respectively. In the illustration in the next paragraph, for instance, we have based  $\pi(\cdot)$  and  $\omega(\cdot)$  on a logistic and linear additive model, respectively, using smoothing splines. The ability to use data-adaptive procedures, makes it more plausible to reason under the assumption that  $\hat{\pi}(\cdot)$  converges to  $\pi(\cdot)$ , which we will make. In that case, the above estimator has been shown (Vansteelandt & Daniel, 2014) to converge to the weighted contrast

$$\frac{E[\pi(L)\{1 - \pi(L)\}\{E(Y|A = 1, L) - E(Y|A = 0, L)\}]}{E[\pi(L)\{1 - \pi(L)\}]}, \quad (3)$$

of the conditional outcome mean at  $A = 1$  versus  $A = 0$ , even when model (1) is misspecified, for example, because  $A$  and  $L$  interact in their association with outcome.

It follows from the above reasoning that the E-estimator, as opposed to the ordinary least squares estimator, is not crucially relying on the restrictions imposed by the outcome model: it returns a meaningful estimand that is directly informative about the conditional association between  $A$  and  $L$ , even when model (1) is misspecified. Even so, caution is warranted as the calculation of standard errors and confidence intervals may still invoke the restrictions of model (1), thereby resulting in overly optimistic inferences about the conditional association between  $A$  and  $Y$ , given  $L$ . This is indeed the case. Standard inference is based on standard errors estimated as 1 over root- $n$  times the sample standard deviation of the so-called (estimated) influence function of  $\hat{\beta}$  under model (1):

$$\frac{\{A_i - \hat{\pi}(L_i)\} \{Y_i - \hat{\beta} A_i - \hat{\omega}(L_i)\}}{n^{-1/2} \sum_{i=1}^n \{A_i - \hat{\pi}(L_i)\} A_i}$$

(Robins et al., 1992). This is valid when model (1) is correctly specified, but ignores that when it is misspecified, then different choices of  $\pi(L)$  in Equation (3) return estimands of a possibly different magnitude. This explains why excess variability, not expressed by the standard deviation of these influence functions, may be observed when repeated samples deliver different estimates of  $\pi(L)$ ; Buja et al. (2019a) make a related remark that such excess variability may lead to differences between fixed- versus random-covariate designs. More formally, as we will see in Section 5, under model misspecification  $\hat{\pi}(L_i)$  contributes to the first-order bias of the E-estimator. This is especially worrying when  $\hat{\pi}(L_i)$  converges (in terms of root mean squared error) at a rate slower than  $n^{-1/2}$ . This may well be the case when smoothing splines are used, and is such that  $\hat{\pi}(L_i)$

will then dominate the behaviour of the E-estimator. Accommodating this can be a daunting task when the behaviour of  $\hat{\pi}(\cdot)$  over repeated samples is ill understood (e.g. because of being based on smoothing splines).

In a simulation study under the above data-generating mechanisms, we found the empirical standard deviation of the E-estimator to be 16.7% larger than estimated, resulting in 87.3% coverage of 95% confidence intervals for Equation (3), despite the lack of bias in  $\hat{\beta}$ . In contrast, the nonparametric approach that we will develop later in this article, resulted in estimators with similar bias, and empirical standard deviation of the E-estimator being only 3.0% larger than estimated (and being only 2.6% larger than that of the E-estimator), resulting in 94.9% coverage of 95% confidence intervals for Equation (3), despite the small sample size ( $n = 50$ ).

### 3 | MAIN EFFECT ESTIMANDS

Suppose that interest lies in the association between a possibly continuous variable or exposure  $A$  and an arbitrary outcome  $Y$ , conditional on measured variables  $L$ . One logical starting point would be the generalised partially linear model

$$g\{E(Y|A, L)\} = \beta A + \omega(L), \quad (4)$$

where  $g(\cdot)$  is a known link function and  $\beta$  and  $\omega(L)$  are unknown. This model choice reflects the fact that in many regression analyses only a small subset of the parameters are of key scientific interest, and an analyst may prefer to be agnostic about the nuisance parameters. Model (4) assumes a linear association as well as the absence of  $A$ - $L$  interactions (on the scale of the link function). It does so for reasons of parsimony, for example, because we may want to summarise the association between  $A$  and  $Y$  into a single number, but not necessarily because it reflects the ground truth. The general question, which we will work out in this paper, is then how to develop inference for  $\beta$  in a way that does not rely on these assumptions.

The starting point of such analysis is to come up with an estimand that is meaningful when the above model does not hold, but reduces to  $\beta$  when the model holds; this then subsequently allows for nonparametric inference to be developed for that estimand. One relatively simple and generic strategy would be to define the estimand as a ‘projection’ of the actual data distribution onto the (semiparametric) model, such as the maximiser of the population expectation of the loglikelihood or some weighted least squares projection (e.g. Buja et al., 2019b; Kennedy et al., 2019; van der Laan & Rose, 2011; Neugebauer & van der Laan, 2007). This suggestion is useful, and we will effectively build on it, but it may deliver estimands that are complicated to interpret. It is moreover vague as there will often be infinitely many such projection estimands. Indeed, each consistent estimator under the (semi)parametric model maps into a projection estimand, being defined as its probability limit under the nonparametric model.

This calls for guidance concerning the choice of estimand in practice. In our development below, we will use three criteria for choosing an estimand. First, when the parametric assumptions hold, it should reduce to the target parameter of interest, in this case the parameter  $\beta$  indexing (4), to assure that the proposal does not hinder a familiar interpretation of the final result. Second, it should be generic, in the sense of being well defined regardless of whether  $A$  is continuous or discrete. Indeed, the fact that parametric methods can flexibly incorporate any type of regressor no doubt contributes to their continuing appeal. It should also be generic in the sense that its efficient influence function should not demand the modelling of a (conditional)



density, as flexible statistical or machine learning techniques are currently not well-adapted to density estimation, and moreover, density estimators may be slowly converging. This criterion distinguishes our development from related work in the causal inference literature, where focus is usually (though not exclusively) given to binary exposures (and effect modifiers). Third, the estimand must capture what one is aiming for (e.g. a conditional association), which was not the case for ordinary least squares in Section 2. This is for instance satisfied when it equals some  $L$ -dependent weighted average of the estimand one would choose to report for a subset of individuals with given  $L$  (e.g. of the average outcome difference between subjects with  $A = 1$  versus  $A = 0$  and the same level of  $L$ ), but is not guaranteed by all projection estimands (see the discussion section).

To distinguish assumptions aimed at parsimony from other, more substantive assumptions, let us start by assuming that the main difficulty of the problem had already been solved. Suppose in particular we already knew  $E(Y|A = a, L)$  for all levels  $a$  in the support of  $A$  and all covariate levels  $L$  over the support of  $L$ . Then we would generally not be interested in reporting exactly how  $E(Y|A = a, L)$  changes over  $a$  and  $L$ . We would content ourselves with a parsimonious summary of the exposure effect. At each level of  $L$ , a useful summary would be the population least squares projection of  $g\{E(Y|A, L)\}$  onto  $A$ , given  $L$ . This reduces to

$$g\{E(Y|A = 1, L)\} - g\{E(Y|A = 0, L)\},$$

when  $A$  is dichotomous (coded 0 or 1). This is clearly capturing a summary of the conditional association between  $A$  and  $Y$ , given  $L$ , regardless of whether some model holds. This  $L$ -specific estimand can next be summarised across levels of  $L$  by taking a weighted average with weights given by

$$\frac{\text{Var}(A|L)}{E\{\text{Var}(A|L)\}};$$

this choice of weights will be motivated later in this section. For dichotomous  $A$ , this delivers the estimand

$$\frac{E(\pi(L)\{1 - \pi(L)\}[g\{E(Y|A = 1, L)\} - g\{E(Y|A = 0, L)\}])}{E[\pi(L)\{1 - \pi(L)\}]}$$

More generally, it gives rise to the estimand

$$\frac{E(\text{Cov}[A, g\{E(Y|A, L)\}|L])}{E\{\text{Var}(A|L)\}}, \quad (5)$$

which reduces to  $\beta$  under model (4), but remains unambiguously defined when this model is misspecified. It will therefore enable us to do inference for  $\beta$  in model (4) without relying on this model restriction. Interpretation of  $\beta$  can still be done in the familiar way, relating to model (4), but with the additional assurance that it continues to represent a summary of the conditional association between  $A$  and  $Y$ , given  $L$ , when that model is misspecified. Such assurance is not attained for standard maximum likelihood estimators, for instance, as we saw in Section 2. We note that the interpretation of our proposed estimand may be more complicated when the  $L$ -specific estimand varies dramatically over levels of  $L$ ; however, other summary measures would also then need to be interpreted with care. Summary measures remain of interest in statistics with the aim to



provide insight, as they may represent all that one can realistically infer with reasonable precision in the face of the curse of dimensionality.

The estimand (5) with  $g(\cdot)$  the identity link has been studied by a number of authors, e.g. Robins et al. (2008), Newey and Robins (2018) and Whitney et al. (2019). We will here extend inference for it to arbitrary link functions. Such extension is non-trivial, if one considers the major difficulties that have been experienced in drawing inference for  $\beta$  under the partially linear logistic model (Tan, 2019; Tchetgen Tchetgen et al., 2010), which have resulted in elegant, but complex proposals that require the modelling of the conditional density or mean of the exposure, given outcome and covariates; relying on such models is arguably less desirable when information about the conditional density of the exposure, given covariates but not outcome, is a priori available (as in randomised experiments, for instance). These complications will be avoided with our choice of estimand (5), which also reduces to  $\beta$  under model (4) with  $g(\cdot)$  the logit link, for which we develop nonparametric inference in Section 5. This extension is moreover important since the probability limits of popular estimators of parameters indexing non-linear models have no simple closed-form representation (unlike the case for the OLS estimator in Section 2), thus rendering their behaviour poorly understood when the model restrictions fail to hold. In particular, estimators for  $\beta$  based on the semiparametric efficient score under the logistic partially linear model will generally fail to converge to (5). We emphasise moreover that our estimand does not require knowing the ‘true’ link function under which the data was generated, since it is defined nonparametrically. Standard advice for fitting generalised linear models is that a link should be chosen that provides a scale where linearity/additivity of the effects of  $A$  and  $L$  is at least plausible. To maintain the connection between our estimand and the parameter in a semiparametric generalised linear model (4), following such advice appears reasonable, although the identity link may yield the simplest interpretation.

When the exposure is dichotomous (taking values 0 and 1),  $g(\cdot)$  is the identity link and moreover  $L$  is sufficient to adjust for confounding (in the sense that  $A$  is independent of the counterfactual outcome  $Y^a$  to exposure level  $a$ , given  $L$ ), then (5) reduces to

$$\frac{E[\pi(L)\{1 - \pi(L)\}(Y^1 - Y^0)]}{E[\pi(L)\{1 - \pi(L)\}]} \quad (6)$$

This effect, which was also considered in Crump et al. (2006) and Vansteelandt and Daniel (2014), gives highest weight to covariate regions where both treated and untreated subjects are found. It expresses the exposure effect that would be observed in a randomised experiment where the chance of recruitment is proportional to both the probability of being treated as well as the probability of being untreated. In that case, subjects with a 10% chance of receiving treatment (or no treatment) are roughly 10 times more likely to be recruited than subjects with a 1% chance of receiving treatment (or no treatment), while subjects whose chance of receiving treatment lies between 25% and 75% are nearly equally likely to be recruited (their chance of recruitment deviates at most 33% in relative terms). Although such recruitment probabilities are not readily applied in a real-life setting, the resulting effect may well approximate that which would be found in a real-life randomised experiment, where the eligibility criteria would exclude patients who are unlikely to receive treatment or no treatment in practice. Regarding the optimality properties of this estimand, Crump et al. (2006) consider the class of weighted sample average treatment effects  $\sum_{i=1}^n w(L_i)(Y_i^1 - Y_i^0) / \sum_{i=1}^n w(L_i)$  where  $w(L)$  is a (known) weight. They show that, under homoscedasticity, the choice  $w(L) = \pi(L)\{1 - \pi(L)\}$  delivers the parameter that can be estimated with the greatest precision across the entire class.

The estimand (5) thus generalises the propensity-overlap-weighted effects to more general exposures and arbitrary link functions. Such generalisation becomes essential when the exposure is continuous, in view of the need to summarise the (now high-dimensional) exposure effect. For binary exposures, an alternative approach which prevents excessive extrapolations would be to consider overlap-weighted effects on other, non-additive scales, for example,

$$\frac{E[\pi(L)\{1 - \pi(L)\}Y^1]}{E[\pi(L)\{1 - \pi(L)\}Y^0]}$$

(Vansteelandt & Daniel, 2014). Such estimands directly target marginal causal effects, as opposed to taking a weighted average of conditional causal effects. They may thus be easier to interpret than (5). However, they do not easily generalise to arbitrary exposures. Moreover, they do not generally reduce to parameters indexing a well-understood generalised linear model, making them arguably more difficult to communicate.

#### 4 | EFFECT MODIFICATION ESTIMANDS

Suppose next that interest lies in the interaction between two possibly continuous variables  $A_1$  and  $A_2$  in their association with a continuous outcome  $Y$ , conditional on measured variables  $L$ . One logical starting point is the generalised partially linear interaction model (Vansteelandt et al., 2008)

$$g\{E(Y|A_1, A_2, L)\} = \omega_1(A_1, L) + \omega_2(A_2, L) + \beta A_1 A_2, \quad (7)$$

where  $\beta$ ,  $\omega_1(A_1, L)$  and  $\omega_2(A_2, L)$  are unknown. The construction of a generic estimand that reduces to  $\beta$  when model (7) is correctly specified, turns out to be a non-trivial task. We are not aware of existing estimands for interaction parameters that satisfy the criteria in Section 3; even if we accept estimands whose efficient influence function requires modelling a density, current proposals are limited to binary  $A_1$  and  $A_2$  (van der Laan & Rose, 2011).

In this paper, we propose to work with the following estimand:

$$\frac{E[\Pi(A_1 A_2)g\{E(Y|A_1, A_2, L)\}]}{E[\Pi(A_1 A_2)^2]}, \quad (8)$$

where  $\Pi(\cdot)$  is an orthogonal projection operator (w.r.t. the covariance inner product), which projects an arbitrary function of  $(A_1, A_2, L)$  onto the space of functions of  $(A_1, A_2, L)$  with mean zero, conditional on  $A_1, L$  as well as conditional on  $A_2, L$ . Such projection eliminates from  $g\{E(Y|A_1, A_2, L)\}$  all main effects of  $A_1$  and  $L$  (as well as their (additive) interactions) and all main effects of  $A_2$  and  $L$  (as well as their (additive) interactions), thus leaving only its dependence on functions of both  $A_1$  and  $A_2$  (and  $L$ ) that cannot be additively separated into functions of  $(A_1, L)$  or  $(A_2, L)$ ; such functions define additive interactions between  $A_1$  and  $A_2$  on the scale of the link function  $g(\cdot)$ . It follows that (8) reduces to  $\beta$  when model (7) is correctly specified. However, a key advantage in pre-specifying such an estimand (relative to standard inference for interactions) is that it continues to capture the interaction between both exposures in their association with outcome, even when this model is misspecified. This is best understood for dichotomous exposures. From the results in Vansteelandt et al. (2008), it then follows that

$$\Pi(A_1 A_2) = \frac{w(L)}{P(A_1, A_2 | L)} \{I(A_1 = A_2) - I(A_1 \neq A_2)\}$$

with

$$w(L) = \left\{ \frac{1}{\pi_{11}(L)} + \frac{1}{\pi_{10}(L)} + \frac{1}{\pi_{01}(L)} + \frac{1}{\pi_{00}(L)} \right\}^{-1},$$

where  $\pi_{a_1 a_2}(L) \equiv P(A_1 = a_1, A_2 = a_2 | L)$  for  $a_1, a_2 = 0, 1$ . With this definition, it can now be shown that (8) reduces to a weighted average of  $L$ -conditional interactions. Indeed, for dichotomous exposures we can always write

$$\begin{aligned} g\{E(Y | A_1, A_2, L)\} &= c_0(L) + c_1(L)A_1 + c_2(L)A_2 \\ &\quad + \{\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L)\} A_1 A_2, \end{aligned}$$

for certain functions  $c_j(L)$ ,  $j = 1, 2, 3$  and  $\mu_{a_1 a_2}(L) \equiv g\{E(Y | A_1 = a_1, A_2 = a_2, L)\}$ . Here,

$$\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L),$$

captures the interaction between both exposures in their association (on the scale of the link function) with outcome, at the considered level of  $L$ . This and the fact that  $c_0(L) + c_1(L)A_1 + c_2(L)A_2$  is orthogonal (w.r.t. the covariance inner product) to  $\Pi(A_1 A_2)$  implies that the estimand (8) reduces to

$$\frac{E[w(L) \{\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L)\}]}{E\{w(L)\}}.$$

Here, the weights  $w(L)$  naturally generalise the propensity-overlap-weights

$$\pi(L)\{1 - \pi(L)\} = \left\{ \frac{1}{\pi(L)} + \frac{1}{1 - \pi(L)} \right\}^{-1},$$

to the setting of interactions between two dichotomous exposures. They assign highest weight to subjects for whom each exposure combination is sufficiently likely, so as to avoid extrapolation towards covariate strata that carry little or no information about interaction. In particular, they down-weight those strata  $L$  in which at least one of the four possible realisations of  $(A_1, A_2)$  is unlikely to be observed. When  $L$  is sufficient to adjust for confounding for the effect of both exposures (in the sense that  $(A_1, A_2)$  is independent of the counterfactual outcome  $Y^{a_1 a_2}$  to exposure  $(a_1, a_2)$ , given  $L$ ) and  $g(\cdot)$  is the identity link, then estimand (8) can also be written as

$$\frac{E\{w(L) (Y^{11} - Y^{10} - Y^{01} + Y^{00})\}}{E\{w(L)\}}. \quad (9)$$

Consider next the special case where  $A_1$  and  $A_2$  are conditionally independent, given  $L$ . This is relevant in settings where  $A_1$  or  $A_2$  is under the control of the investigator (such that  $A_1$  is independent of  $A_2$  is known to hold by design); for example in summarising how the effect of a randomised treatment  $A_1$  is modified by a continuous covariate  $A_2$ . It is moreover relevant in gene-environment interaction studies (where it is usually assumed that genetic and

environmental factors are independent in the population). Then it further follows from Vansteelandt et al. (2008) that

$$\Pi(A_1 A_2) = \{A_1 - E(A_1|L)\} \{A_2 - E(A_2|L)\},$$

regardless of whether the exposures are dichotomous or not. In that case, the estimand (8) can also be written as

$$\frac{E \left[ \{A_1 - E(A_1|L)\} \{A_2 - E(A_2|L)\} g\{E(Y|A_1, A_2, L)\} \right]}{E \left[ \{A_1 - E(A_1|L)\}^2 \{A_2 - E(A_2|L)\}^2 \right]}. \quad (10)$$

When  $A_1$  and  $A_2$  are dichotomous, this simplifies further to

$$\frac{E \left[ \pi_1(L) \{1 - \pi_1(L)\} \pi_2(L) \{1 - \pi_2(L)\} \{\mu_{11}(L) + \mu_{00}(L) - \mu_{10}(L) - \mu_{01}(L)\} \right]}{E \left[ \pi_1(L) \{1 - \pi_1(L)\} \pi_2(L) \{1 - \pi_2(L)\} \right]}, \quad (11)$$

where  $\pi_1(L) = P(A_1 = 1|L)$  and  $\pi_2(L) = P(A_2 = 1|L)$ .

In the more general case, the projection operator is not obtainable in closed-form but can be obtained via the alternating conditional expectations (ACE) algorithm (Bickel et al., 1993). This involves first taking the difference  $U_1$  between  $A_1 A_2$  and its conditional expectation given  $A_1$  and  $L$ , next taking the difference  $U_2$  between  $U_1$  and its conditional expectation given  $A_2$  and  $L$ , next taking the difference  $U_3$  between  $U_2$  and its conditional expectation given  $A_1$  and  $L$ , and so on ..., eventually delivering the projection  $U_\infty$ . Importantly, this algorithm does not demand knowledge of the entire joint density of both exposures, conditional on  $L$ , and moreover avoids inverse weighting by such density. This is essential for enabling an inference that is generic (e.g. can be used for continuous exposures), and has been a key challenge in proposing a generic estimand such as (8).

## 5 | NONPARAMETRIC INFERENCE

In the previous sections, we have shown how modelling assumptions can be invoked to summarise the (conditional) association between two variables, which may itself be high-dimensional, or the extent to which two variables interact in their association with an outcome. To prevent that these convenience assumptions are used as a ground truth, we next develop inference for the resulting estimands under a nonparametric model.

### 5.1 | Main effect estimands

#### 5.1.1 | The plug-in estimator

A natural estimator of the main effect estimand  $\beta$ , given by Equation (5), is

$$\frac{\sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\} g\{\hat{E}(Y_i|A_i, L_i)\}}{\sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\} A_i}. \quad (12)$$

We call it a ‘plug-in’ estimator, as it equals the sample analogue of Equation (5) with estimators  $\hat{E}(A|L)$  and  $\hat{E}(Y|A, L)$  of the unknown conditional expectations ‘plugged in’. In the spirit of being ‘assumption-free’ (or at least, assumption-lean) it is natural to learn these conditional expectations without pre-specification of parametric models. One could therefore adopt variable/model selection procedures, or use traditional nonparametric estimators (e.g. kernel methods, sieve estimators, regression trees) or even machine learning approaches (random forests, neural networks, support vector machines) which are particularly effective when the dimension of the covariates is large. Machine learning techniques learn a (potentially very complex) ‘model’ from the data, while using regularisation (in combination with cross-validation) to minimise issues of overfitting and optimise out-of-sample predictive performance. The analyst does not need choose between different estimators now available in statistical software; ensemble learners, such as the Super Learner (van der Laan et al., 2007), aim to take the optimal weighted combination of candidate (parametric and nonparametric) estimators.

Traditionally, statisticians have been hesitant to routinely incorporate machine learning when analysing data. This is in part because the tuning parameters used to control the degree of regularisation in the data-adaptive estimators  $\hat{E}(A|L)$  and  $\hat{E}(Y|A, L)$  are typically chosen to balance bias and variance in a way that is optimal for prediction purposes. Unfortunately, this choice is usually *suboptimal* for estimation of the target parameter; the ‘plug-in’ estimator of  $\beta$  given in Equation (12) can inherit the potentially large biases from  $\hat{E}(A|L)$  and  $\hat{E}(Y|A, L)$ . The consequence is that the bias of the naive estimator may be of the order  $n^{-1/2}$  or larger, and hence the use of standard confidence intervals is not justified. A further issue is that even if parametric-rate confidence intervals could be constructed, it is unclear how one would account for the uncertainty in the estimation of the nuisance parameters, given that these may follow a complex distribution.

### 5.1.2 | The efficient influence function

To overcome the problems associated with plug-in estimators, we will develop inference for the estimand under a nonparametric model based on its so-called efficient influence function (Bickel et al., 1993; Pfanzagl, 1990). Technically, this is mean zero functional of the observed data and the data-generating distribution, which characterises the estimand’s sensitivity to arbitrary (smooth) changes in the data-generating law. The efficient influence function for the proposed estimand is given below.

**Theorem 1** *Under the nonparametric model, the main effect estimand  $\beta$ , defined by Equation (5), has efficient influence function*

$$\frac{\{A - E(A|L)\}[\mu(Y, A, L) - \beta\{A - E(A|L)\}]}{E\{[A - E(A|L)]^2\}} \quad (13)$$

where  $g'(x) = \partial g(x)/\partial x$  and

$$\mu(Y, A, L) = g'\{E(Y|A, L)\}\{Y - E(Y|A, L)\} + g\{E(Y|A, L)\} - E[g\{E(Y|A, L)\}|L].$$

The proof of this and all other results is given in Section 1 of the Supplemental Materials.

If the conditional expectations indexing the efficient influence function were known, then it would follow from its mean zero property that a consistent estimator  $\tilde{\beta}$  of  $\beta$  could be obtained

as the value of  $\beta$  that makes the sample average of the influence functions zero. The resulting estimator's asymptotic distribution would be governed by this influence function in the sense that

$$\sqrt{n}(\tilde{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\{A_i - E(A_i|L_i)\} [\mu(Y_i, A_i, L_i) - \beta \{A_i - E(A_i|L_i)\}]}{E[\{A - E(A|L)\}^2]} + o_p(1). \quad (14)$$

The fact that the difference between the estimator and the truth can be approximated by the sample average of a mean-zero random variable implies that  $\tilde{\beta}$  is asymptotically normally distributed with bias that shrinks to zero faster than the standard error, and with a variance that can be estimated as the sample variance of the efficient influence function (where population expectations and the value of  $\beta$  can be substituted by consistent estimates).

The fact that the efficient influence function involves unknown conditional expectations, makes the estimator  $\tilde{\beta}$  suggested in the previous paragraph infeasible. As in the previous section, we will therefore substitute these by consistent estimators and denote the resulting estimator  $\hat{\beta}$ . The power of basing the estimator on the efficient influence function is that it behaves the same asymptotically whether it be based on known conditional expectations or consistent estimators thereof, provided that these converge sufficiently fast in a relatively weak sense (made specific in the following theorem). Throughout this section,  $\mathbb{P}_n$  denotes the empirical measure (i.e. sample average) and for a function  $f(O)$  of the data  $O$  we use the notation  $\mathbb{P}\{f(O)\} = \int f(O)\mathbb{P}(O)dO$  where  $\mathbb{P}(O)$  denotes the density of the data; for an estimator  $\hat{f}$ ,  $\mathbb{P}\{\hat{f}(O)\}$  averages over  $O$  but not  $\hat{f}$ .

**Theorem 2** *Let  $\hat{\beta}$  refer to the proposed estimator of  $\beta$  based on estimators  $\hat{E}(A|L)$  and  $\hat{\mu}(Y, A, L)$  which are consistent for  $E(A|L)$  and  $\mu(Y, A, L)$ , respectively (see details in the Appendix). Suppose that the weak positivity assumptions at both the population and sample level hold that  $\mathbb{P}[\{A - E(A|L)\}^2] > \sigma$ ,  $\mathbb{P}_n[\{A - E(A|L)\}^2] > \sigma$  and  $\mathbb{P}_n[\{A - \hat{E}(A|L)\}^2] > \sigma$  for some  $\sigma > 0$ . Suppose furthermore that at least one of the following two conditions hold:*

1. (Sample-splitting)  $\hat{E}(A|L)$  and  $\hat{\mu}(Y, A, L)$  are obtained from a sample independent from the one used to construct  $\hat{\beta}$ .
2. (Donsker condition) The quantity

$$\frac{\{A - \hat{E}(A|L)\}[\hat{\mu}(Y, A, L) - \hat{\beta}\{A - \hat{E}(A|L)\}]}{\mathbb{P}_n[\{A - \hat{E}(A|L)\}^2]}$$

*falls within a  $\mathbb{P}$ -Donsker class with probability approaching 1.*

Finally, assume that  $A - \hat{E}(A|L) = O_p(1)$  and that sufficient rates of convergence are attained so that the following terms are  $o_p(n^{-1/2})$ :

$$\begin{aligned} & \mathbb{P}\left[\{E(Y|A, L) - \hat{E}(Y|A, L)\}^2\right], \\ & \mathbb{P}\left[\{E(A|L) - \hat{E}(A|L)\}^2\right], \\ & \mathbb{P}\left[\{E(A|L) - \hat{E}(A|L)\}^2\right]^{1/2} \mathbb{P}\left\{\left(E[g\{E(Y|A, L)\}|L] - \hat{E}[g\{\hat{E}(Y|A, L)\}|L]\right)^2\right\}^{1/2}, \end{aligned}$$

Then it follows that (14) holds with  $\hat{\beta}$  in lieu of  $\tilde{\beta}$ .

A detailed discussion of the above assumptions is saved for later on in this section. A consequence of this result is that the variance of  $\hat{\beta}$  can be estimated as previously suggested, namely

as 1 over  $n$  times the sample variance of the efficient influence functions, as if these conditional expectations were given. It implies in particular that the uncertainty that the estimators of the conditional expectations add to the analysis can be ignored when drawing inference about  $\beta$ , even when these are based on variable selection or machine learning procedures, whose uncertainty is difficult to quantify.

We can thus obtain an estimator and confidence interval via the simple recipe below:

1. Obtain estimates  $\hat{E}(A|L)$  and  $\hat{E}(Y|A, L)$ , e.g. using machine learning.
2. If  $A$  is binary, estimate  $E[g\{E(Y|A, L)\}|L]$  as

$$\hat{E}[g\{\hat{E}(Y|A, L)\}|L] = g\{\hat{E}(Y|A = 1, L)\}\hat{E}(A|L) + g\{\hat{E}(Y|A = 0, L)\}\{1 - \hat{E}(A|L)\}$$

otherwise, use an additional data-adaptive fit (with  $g\{\hat{E}(Y|A, L)\}$  as outcome).

3. Fit a linear regression of

$$\begin{aligned} \hat{\mu}(Y, A, L) = & g'\{\hat{E}(Y|A, L)\}\{Y - \hat{E}(Y|A, L)\} \\ & + g\{\hat{E}(Y|A, L)\} - \hat{E}[g\{\hat{E}(Y|A, L)\}|L] \end{aligned}$$

on the sole predictor  $A - \hat{E}(A|L)$  (without an intercept) using OLS in order to obtain an estimate  $\hat{\beta}$  of  $\beta$ .

The variance of  $\hat{\beta}$  can be consistently estimated as

$$\hat{V}(\hat{\beta}) = \frac{n^{-2} \sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\}^2 [\hat{\mu}(Y_i, A_i, L_i) - \hat{\beta} \{A_i - \hat{E}(A_i|L_i)\}]^2}{\left[ n^{-1} \sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\}^2 \right]^2}.$$

It is readily obtained by requesting that the software provide a sandwich estimator in step (c).

A confidence interval can be constructed as  $\hat{\beta} \pm 1.96 \sqrt{\hat{V}(\hat{\beta})}$ .

The rate conditions required in Theorem 2 will hold if all nuisance parameters are consistently estimated at a rate faster than  $n^{1/4}$ ; under certain smoothness/sparsity assumptions, these are attainable for many data-adaptive methods (see Chernozhukov et al. (2018) for a summary). The Donsker condition, which restricts the complexity of the estimators involved, is unlikely to be satisfied for very flexible machine learning methods. A simple solution is to use sample-splitting; split the data in half, estimate the nuisance parameters in the ‘training’ split, and perform inference on  $\beta$  in the ‘validation’ sample. This has a disadvantage of halving the sample size. However, efficiency can be asymptotically recovered via cross-fitting (Chernozhukov et al., 2018; Zheng & van der Laan, 2011); for example, one can reverse the training and validation samples, construct a second estimate of  $\beta$  and average the pair. Confidence intervals can be constructed by combining the estimated influence functions across the different splits, replacing  $\beta$  with the averaged rather than split-specific estimate. As before, one can then estimate the variance of cross-fit estimator of  $\beta$  as 1 over  $n$  times the sample variance of the (estimated) influence functions.

The combination of efficient influence function-based estimators with cross-fitting facilitates the use of machine learning to estimate parts of the data distribution of no scientific interest. These important results have only been highlighted relatively recently (Chernozhukov et al., 2018;



Zheng & van der Laan, 2011), and many open questions remain. Firstly, there is yet to be firm guidance on the number of splits to use in the cross-fitting. Moreover, since the machine learning methods typically perform better with more data, it may be that no splitting can sometimes yield estimators of the target parameter with smaller bias (though potentially more biased standard errors) compared to cross-fitting. At the other extreme, due to the similarity of our estimator to that of Robinson (1988), it may be possible to obtain much sharper results on the nuisance estimators by using a more specific variant of cross-fitting in combination with so-called ‘under-smoothing’ (Newey & Robins, 2018). This is left to future work. For now, if cross-fitting is adopted, we recommend 10-fold cross-fitting, each time using nine tenths as training sample and the remainder as validation sample.

### 5.1.3 | Illustration - inference under the partially linear model

We return to the case study in Section 2. So long as the partially linear model (1) holds, it turns out that there are several different ways of constructing estimators of  $\beta$  that are desensitised to ‘plug-in’ bias of machine learners. Chernozhukov et al. (2018) propose using either the E-estimator (2) described in Section 2 or the ‘partialling out’ estimator

$$\frac{\sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\} \{Y_i - \hat{E}(Y_i|L_i)\}}{\sum_{i=1}^n \{A_i - \hat{E}(A_i|L_i)\}^2}, \quad (15)$$

(Robinson, 1988), where all nuisance parameters are estimated via machine learning. So long as the semiparametric model restriction holds, under standard conditions both estimation approaches discussed in the previous paragraph are first-order equivalent. The E-estimator has an influence function

$$\frac{\{A - E(A|L)\} \{Y - \beta A - E(Y|A = 0, L)\}}{E \left[ \{A - E(A|L)\}^2 \right]}$$

that coincides with the influence function

$$\frac{\{A - E(A|L)\} [Y - E(Y|L) - \beta \{A - E(A|L)\}]}{E \left[ \{A - E(A|L)\}^2 \right]}$$

for the ‘partialling out’ estimator when model (1) holds. The latter reduces to (13) for the identity link, given in Theorem 1; indeed, the estimators proposed in the previous subsection generalise the ‘partialling out’ estimator to arbitrary link functions. Further, assuming the residual variance of  $Y$  conditional on  $A$  and  $L$  is a constant  $\sigma^2$ , both estimators have an asymptotic variance equal to the semiparametric efficiency bound  $\sigma^2 / E\{\text{var}(A|L)\}$ . The asymptotic bias of both approaches depends in part on the product of two errors - either

$$\{E(A|L) - \hat{E}(A|L)\} \{E(Y|A = 0, L) - \hat{E}(Y|A = 0, L)\}$$

for the E-estimator or

$$\{E(A|L) - \hat{E}(A|L)\} [E(Y|L) - \hat{E}(Y|L) - \beta \{E(A|L) - \hat{E}(A|L)\}] \quad (16)$$

for the ‘partialling out’ estimator. As long as each estimator converges to the truth, then the product of two errors will tend to shrink at least as fast (and usually much faster) than an individual error.

However, the situation is quite different when restriction (1) fails (Whitney et al., 2019). The asymptotic bias of the E-estimator, relative to estimand (5), is now proportional to

$$E[\{E(A|L) - \tilde{E}(A|L)\}\{E(Y - \beta A|L) - \tilde{E}(Y|A = 0, L)\}],$$

where  $\tilde{E}(A|L)$  is the probability limit of  $\hat{E}(A|L)$  and  $\tilde{E}(Y|A = 0, L)$  is the probability limit of  $\hat{E}(Y|A = 0, L)$ ; note that  $E(Y|A = 0, L) = E(Y - \beta A|L)$  under the partially linear model but not otherwise. Because the error  $E(Y - \beta A|L) - \tilde{E}(Y|A = 0, L)$  will no longer shrink to zero, the bias of the E-estimator will be determined by  $E(A|L) - \tilde{E}(A|L)$ . As discussed above, the situation may be much worse for semiparametric estimators in nonlinear models, since the bias w.r.t (5) may now even diverge. By considering (16), it follows that the same issues are not true for the ‘partialling out’ estimator, which makes the sample average of the influence functions for the estimand (5) evaluated at the machine learning predictions equal to zero. This highlights the benefits of estimation using the influence function obtained under a nonparametric model; it incorporates an implicit bias-correction, as the bias of the estimator of the target parameter is usually smaller in magnitude than that of the first stage estimators. Moreover, this property is not dependent on any semiparametric modelling assumptions.

Note also that when model (1) is misspecified, each change of  $\pi(L)$  also changes the estimand targeted by the E-estimator. In particular, different estimates of the propensity score may then be viewed as targeting different effect estimands. The resulting excess variability is not acknowledged when basing inference on the influence function of the E-estimator, as this is assuming model (1) to be correctly specified. This was indeed what was observed in the simulation study described in Section 2. As Buja et al. (2019c) note, for certain choices of nuisance parameter estimators (specifically, series methods or twicing kernels) the E-estimator and the proposed influence function-based estimator can exactly coincide. However, since we wish to work in greater generality, and in the following section consider arbitrary machine learners for the nuisances, we do not consider this subtlety any further.

## 5.2 | Effect modification estimands

The following theorem gives the efficient influence function for the effect modification estimand  $\beta$ , given by Equation (8), under the nonparametric model.

**Theorem 3** *Under the nonparametric model, the effect modification estimand  $\beta$ , defined by Equation (11), has efficient influence function*

$$\frac{\Pi(A_1 A_2)}{E\{\Pi(A_1 A_2)^2\}} \{\mu(Y, A_1, A_2, L) - \beta \Pi(A_1 A_2)\}$$

where

$$\mu(Y, A_1, A_2, L) \equiv g' \{E(Y|A_1, A_2, L)\} \{Y - E(Y|A_1, A_2, L)\} + \Pi \{g \{E(Y|A_1, A_2, L)\}\}.$$

A root- $n$  consistent estimator of  $\beta$  can thus be obtained as

$$\hat{\beta} = \left\{ \frac{1}{n} \sum_{i=1}^n \hat{\Pi}^2(A_{i1}A_{i2}) \right\}^{-1} \frac{1}{n} \sum_{i=1}^n \hat{\Pi}(A_{i1}A_{i2}) \hat{\mu}(Y_i, A_{i1}, A_{i2}, L_i),$$

where

$$\hat{\mu}(Y, A_1, A_2, L) = g' \{ \hat{E}(Y|A_1, A_2, L) \} \{ Y - \hat{E}(Y|A_1, A_2, L) \} + \hat{\Pi} [g \{ \hat{E}(Y|A_1, A_2, L) \}].$$

Here,  $\hat{E}(Y|A_1, A_2, L_i)$  denotes a data-adaptive prediction (e.g. obtained using machine learning or a flexible parametric model building procedure). Furthermore, the projection  $\hat{\Pi}(A_{i1}A_{i2})$  can be obtained via the ACE algorithm (Bickel et al., 1993). This involves first data-adaptively predicting  $A_{i1}A_{i2}$  on the basis of  $A_{i1}$  and  $L_i$  and taking the residuals; next, data-adaptively predict these residuals on the basis of  $A_{i2}$  and  $L_i$  and take the residuals; next, data-adaptively predict these residuals on the basis of  $A_{i1}$  and  $L_i$  and take the residuals; and so forth. This process can be aborted when the variance of the predicted residuals reaches a value very close to zero. To ensure a decreasing variance, we recommend in each step tuning the obtained predictions of the residuals by substituting these by the ordinary least squares prediction of those residuals onto the obtained data-adaptive predictions. The projection  $\hat{\Pi} \{ \hat{E}(Y|A_{i1}, A_{i2}, L_i) \}$  is likewise obtained, starting from  $\hat{E}(Y|A_{i1}, A_{i2}, L_i)$ . The following theorem outlines the necessary conditions on the nuisance parameters, in order to obtain valid inference.

**Theorem 4** Suppose that estimators  $\hat{\Pi}(\cdot)$  and  $\hat{E}(Y|A_1, A_2, L)$  are consistent for  $\Pi(\cdot)$  and  $E(Y|A_1, A_2, L)$ , respectively (see details in the Appendix). Suppose that the weak positivity assumptions at both the population and sample level hold that  $\mathbb{P} \{ \Pi(A_1A_2)^2 \} > \sigma$ ,  $\mathbb{P}_n \{ \Pi(A_1A_2)^2 \} > \sigma$  and  $\mathbb{P}_n \{ \hat{\Pi}(A_1A_2)^2 \} > \sigma$  for some  $\sigma > 0$ . Suppose furthermore that at least one of the following two conditions hold:

1. (Sample-splitting)  $\hat{\Pi}(A_1A_2)$ ,  $\hat{E}(Y|A_1, A_2, L)$  and  $\hat{\Pi}[g\{\hat{E}(Y|A_1, A_2, L)\}]$  are obtained from a sample independent to the one used to construct  $\hat{\beta}$ .
2. (Donsker condition) The quantity

$$\frac{\hat{\Pi}(A_1A_2)}{\mathbb{P}_n \{ \hat{\Pi}(A_1A_2)^2 \}} \{ \hat{\mu}(Y, A_1, A_2, L) - \hat{\beta} \hat{\Pi}(A_1A_2) \}$$

falls within a  $\mathbb{P}$ -Donsker class with probability approaching 1.

Further, assume that  $\hat{\Pi}(A_1A_2) = O_p(1)$  and that sufficient rates of convergence are attained so that the following terms are  $o_p(n^{-1/2})$ :

$$\begin{aligned} & \mathbb{P} \left[ \{ E(Y|A, L) - \hat{E}(Y|A, L) \}^2 \right], \\ & \mathbb{P} \left[ \{ \Pi(A_1A_2) - \hat{\Pi}(A_1A_2) \}^2 \right], \\ & \mathbb{P} \left( \{ \Pi(A_1A_2) - \hat{\Pi}(A_1A_2) \}^2 \right)^{1/2} \mathbb{P} \left( [\Pi[g\{E(Y|A, L)\}] - \hat{\Pi}[g\{\hat{E}(Y|A, L)\}]]^2 \right)^{1/2}, \end{aligned}$$

where, for a random variable  $V$ ,  $\hat{\Pi}^*(V) \equiv V - \hat{\Pi}(V)$ . Then it follows that

$$\sqrt{n}(\hat{\beta} - \beta) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{\Pi(A_{i1}A_{i2})}{E\{\Pi(A_{i1}A_{i2})^2\}} \{\mu(Y_i, A_{i1}, A_{i2}, L_i) - \beta \Pi(A_{i1}A_{i2})\} + o_p(1).$$

The variance of both considered estimators is obtained as 1 over  $n$  times the variance of the corresponding influence function, with conditional expectations substituted by data-adaptive predictions, marginal expectations by sample averages and  $\beta$  by  $\hat{\beta}$ .

## 6 | SIMULATION STUDIES

To provide insight into different aspects of the proposal, we provide results on 4 sets of simulation experiments. In all experiments, we report Monte Carlo bias and standard deviation (SD), as well as standard errors (SE) estimated as 1 over root- $n$  times the sample standard deviation of the estimated influence functions and coverage of corresponding 95% Wald confidence intervals. Throughout, we will refer to the proposal as ‘AL’ for ‘Assumption-Lean’.

### 6.1 | Main effects, binary exposure

In the first experiment, we study inference for the main effect estimand (5) with  $g(\cdot)$  the logit link. The aim of this experiment is to contrast our proposal based on random forest regression with 3 competing estimators. In particular, we considered the maximum likelihood estimator (MLE) of  $\beta$  obtained by fitting the logistic regression model  $\text{logit}\{E(Y|A, L)\} = \beta A + \alpha_0 + \alpha_1^T L$ . We also included two estimators designed for the partially linear logistic model  $\text{logit}\{E(Y|A, L)\} = \beta A + \omega(L)$ ; the first estimator ‘ES’ solves the semiparametric efficient score equations e.g. in Kosorok (2007):

$$0 = \sum_{i=1}^n \left( A_i - \frac{\hat{E} \left[ A_i \hat{E}(Y_i|A_i, L_i) \{1 - \hat{E}(Y_i|A_i, L_i)\} | L_i \right]}{\hat{E} \left[ \hat{E}(Y_i|A_i, L_i) \{1 - \hat{E}(Y_i|A_i, L_i)\} | L_i \right]} \right) \times (Y_i - \text{expit}[\beta A_i + \text{logit}\{\hat{E}(Y_i|A_i = 0, L_i)\}])$$

whereas the second is the simple doubly robust (DR) estimator proposed in Tchetgen Tchetgen (2013), which solves the equations

$$0 = \sum_{i=1}^n \{A_i - \hat{E}(A_i|Y_i = 0, L_i)\} \{Y_i - \hat{E}(Y_i|A_i = 0, L_i)\} \exp(-\beta A_i Y_i).$$

For this purpose, we generated a 10-dimensional covariate  $L \sim N(0, \Sigma)$ , where  $\Sigma$  was (once) randomly generated with variances between 2 and 10 and correlations up to 0.72 in absolute value and then fixed across simulations; and  $A \sim \text{Bern}(\gamma^T L - 0.15L_1^2)$ , where  $\gamma$  is the 10-dimensional unit vector scaled by  $1/\sqrt{40}$  and  $L_k$  is the  $k$ th entry of  $L$ . For generating the outcome  $Y$ , we considered 4 separate settings: (1)  $Y \sim \text{Bern}(\text{expit}(0.3A + \delta^T L_{[1:5]}))$  where  $\delta$  is a 5-dimensional unit vector scaled by  $1/10$ ; (2)  $Y \sim \text{Bern}(\text{expit}(0.3A + \delta^T L_{[1:5]} + 0.1L_1^2))$ ; (3)  $Y \sim \text{Bern}(\text{expit}(L_1(1.5A - 1) + \delta^T L_{[1:5]}))$ ; and (4)  $Y \sim \text{Bern}(\text{expit}(0.1/(1 + \exp(0.1L_3 - 0.1L_2)) + 0.3A/(1 + \exp(-0.1L_2)) +$

$0.5AL_6 + 0.025L_1^2$ ). Only in the first two settings does the partially linear logistic model restriction hold; the fourth setting is especially challenging, in light of the complex functional form of the interaction between  $A$  and  $L$ .

For the ES estimator as well as the proposal, random forests (via the ‘grf’ package described in Athey et al. (2019)) were used to learn  $E(Y|A, L)$  and  $E(A|L)$  and yield predictions. These could then be plugged into the relevant estimating equations via application of the law of total probability. For the DR estimator, random forests were used to learn  $E(A|Y, L)$  so that predictions of  $E(A|Y = 0, L)$  could be obtained, as this reflects how this conditional expectation would likely be estimated in practice using machine learning. In experiments 3 and 4, in order not to privilege the proposed estimand, we chose to report bias and coverage relative to the population limit of the estimator. The latter was approximated by generating 500 datasets with sample size 100,000, using the true conditional expectations for the nuisance functionals where possible, and taking the average of the resulting estimates.

In Table 1, we see that the MLE does not always target an estimand that summarises the conditional association of scientific interest. This is confirmed in experiment 2, where the limit of the MLE is in a different direction to the parameter in the partially linear model, which is especially worrisome. We see that the two semiparametric approaches perform well when the model restriction holds. In experiment 3 and 4 however we see that outside of the model, coverage can sharply decrease as sample size increases. This is the result of excess variability not reflected by standard errors when the model is misspecified. This is particularly the case for the DR estimator, where the bias inherited from the random forests appears to be substantial. Our proposal had better coverage than competing approaches in experiment 3 across the different sample sizes; this is due to both lower bias, and estimated standard errors that at least in large samples more accurately reflect the variability of the estimator. Reassuringly, despite our inferences being assumption-lean, the empirical standard deviations show that this does not come with a loss of precision.

## 6.2 | Effect modification, binary exposure

In a second set of simulation experiments, we considered inference for effect modification estimand (8), with  $g(\cdot)$  the identity link and without making the assumption of conditionally independent exposures. We generated a 10-dimensional covariate  $L \sim N(0, \Sigma)$ , where  $\Sigma$  was (once) randomly generated as before. The exposure was generated as in the previous section, and the outcome as  $Y \sim N(3/(1 + \exp(L_3 - L_2)) + A/(1 + \exp(L_1 - L_2)), 1)$ . This data-generating mechanism is inspired by Nie and Wager (2017), but made more complicated by means of a non-randomised exposure  $A$ .

Our aim was to assess evidence for modification of the effect of  $A$  by  $L_3$ . Since such effect modification is absent, we here studied the performance of different estimation methods w.r.t. their ability to retrieve zero effect modification (thus also giving us a different perspective than in the previous section, where we contrasted each estimator with its limit in experiments 3 and 4). The simulation results in Table 2 demonstrate favourable results for the proposal, based on random forests (via the ‘grf’ package described in Athey et al. (2019)) as compared to OLS based on a linear model that includes all main effects along with the interaction between  $A$  and  $L_3$ . In particular, we observe smaller bias and better coverage at the expense of an increase in standard errors (around 30% larger).

In a second set of simulation experiments, we made the data-generating mechanism even more challenging by changing the outcome model to  $Y \sim N(3/(1 + \exp(L_3 - L_2)) + A/(1 +$

**TABLE 1** Simulation results on main effects: empirical bias (Bias) and standard deviation (SD), sample average of the estimated influence-function based standard errors (SE), and coverage of 95% Wald confidence intervals (Cov). Bias and coverage taken w.r.t. the truth 0.3 in experiments (Exp.) 1 and 2, and w.r.t. the limiting values of each estimator in experiment 3 (0.33 (MLE) 0.43 (ES), 1.00 (DR) and 0.50 (AL)) and in experiment 4 (−0.08 (MLE), 0.19 (ES) 0.37 (DR) and 0.21 (AL))

Exp.	Est.	n = 500				n = 1000				n = 2000			
		Bias	SD	SE	Cov	Bias	SD	SE	Cov	Bias	SD	SE	Cov
1	MLE	0.00	0.21	0.21	95	0.00	0.15	0.15	95	0.00	0.11	0.10	94
	ES	0.04	0.20	0.19	93	0.03	0.15	0.14	92	0.02	0.11	0.10	92
	DR	0.06	0.21	0.23	96	0.05	0.15	0.16	96	0.03	0.11	0.11	95
	AL	0.02	0.19	0.20	95	0.02	0.14	0.14	95	0.01	0.10	0.10	94
2	MLE	−0.59	0.22	0.22	26	−0.59	0.15	0.16	3	−0.59	0.11	0.11	0
	ES	−0.14	0.21	0.20	86	−0.04	0.16	0.14	90	−0.02	0.12	0.10	91
	DR	−0.11	0.22	0.24	90	−0.01	0.17	0.18	96	0.01	0.12	0.13	95
	AL	−0.17	0.20	0.21	89	−0.07	0.15	0.15	92	−0.04	0.11	0.11	93
3	MLE	0.01	0.26	0.23	92	0.01	0.18	0.16	94	0.00	0.13	0.12	94
	ES	0.04	0.28	0.23	88	0.05	0.22	0.18	88	0.02	0.17	0.14	88
	DR	−0.60	0.21	0.19	16	−0.50	0.16	0.14	7	−0.40	0.13	0.10	6
	AL	−0.05	0.28	0.23	88	0.00	0.22	0.19	91	0.01	0.17	0.15	92
4	MLE	−0.01	0.20	0.21	95	0.00	0.15	0.15	95	0.00	0.10	0.10	96
	ES	−0.11	0.20	0.19	91	−0.06	0.16	0.14	90	−0.04	0.12	0.11	91
	DR	−0.29	0.21	0.21	69	−0.25	0.16	0.15	61	−0.22	0.11	0.11	48
	AL	−0.12	0.19	0.20	92	−0.07	0.15	0.14	91	−0.04	0.12	0.11	92

$\exp(L_1 - L_2)) + 5AL_6, 1)$ . The inclusion of an interaction between  $A$  and  $L_6$  now makes it increasingly difficult to demonstrate the absence of effect modification between  $A$  and  $L_3$  (which has a correlation of  $-0.54$  with  $L_6$ ). The simulation results demonstrate drastically favourable results for the proposal with a much smaller bias as well as standard errors (up to 4 times smaller than for OLS), resulting in much better coverage.

To demonstrate the behaviour under conditions where the linear regression model is correctly specified, we additionally generated a continuous exposure  $A \sim N(\gamma^T L, 1)$ , where  $\gamma$  is the  $d$ -dimensional unit vector scaled by  $1/\sqrt{40}$ , and the outcome as  $Y \sim N(\gamma^T L + 5AL_3, 1)$ . Both methods give good performance in this setting, with the proposal not surprisingly delivering larger standard errors (roughly up to 2.5 times larger) in view of the poorer ability of random forest regression to pick up linear trends. Here, better performance can be expected with the use of ensemble learners.

### 6.3 | High-dimensional variable selection, continuous exposure

In a third set of simulation experiments, we considered inference for the main effect estimand (5) with  $g(\cdot)$  the logit link in the presence of high-dimensional covariates using the data-generating mechanism in Belloni et al. (2013). In particular, we generated a 250-dimensional covariate

**TABLE 2** Simulation results on effect modification: empirical bias (Bias) and standard deviation (Emp SD), sample average of the estimated influence-function based standard errors (Mean SE), and coverage of 95% Wald confidence intervals (Cov)

Exp.	Est.	<i>n</i> = 500				<i>n</i> = 1000				<i>n</i> = 2000			
		Bias	SD	SE	Cov	Bias	SD	SE	Cov	Bias	SD	SE	Cov
1	OLS	−0.047	0.051	0.051	84	−0.046	0.037	0.036	76	−0.046	0.027	0.025	55
	AL	−0.034	0.067	0.073	95	−0.016	0.051	0.050	93	−0.015	0.036	0.035	92
2	OLS	−2.92	0.24	0.23	0	−2.92	0.17	0.16	0	−2.92	0.11	0.11	0
	AL	−0.31	0.15	0.16	49	−0.12	0.077	0.085	77	−0.057	0.044	0.052	88
3	OLS	0.00	0.015	0.015	94	0.00	0.010	0.010	95	0.00	0.007	0.007	97
	AL	0.019	0.042	0.043	93	0.013	0.027	0.029	95	0.002	0.018	0.021	97

$L \sim N(0, \Sigma)$ , where  $\Sigma$  is an autoregressive correlation matrix with correlation parameter 0.5. The exposure was normally distributed with mean given by  $\sum_{j=1}^{10} L_j / j$  and unit residual variance. The outcome was dichotomous with mean given by  $\text{expit} \left[ 0.2A + \sum_{j=1}^5 L_j / (2j) + \sum_{j=11}^{15} L_j / \{2(j - 10)\} \right]$ .

We evaluated the performance of the standard lasso and elastic-net estimators under a main effect logistic regression model, as well as the post-lasso (P-lasso) estimator obtained by refitting that model using the selected variables. In each case, the penalty was chosen as the largest value for which the cross-validated prediction error is within 1 standard error of the minimum. We moreover evaluated the proposed assumption-lean procedure (AL) based on these fitting strategies for both the outcome and exposure, assuming that these obey main effect logistic and linear models, respectively. We finally also included a plug-in estimator (SL) and the proposed estimator (AL SL) based on SuperLearner fits for  $E(Y|A, L)$ , for  $E[g\{E(Y|A, L)\}]$  and for  $E(A|L)$ . The SuperLearner library included two lasso procedures and two elastic-net procedures, using penalties equal to either the above suggested penalty or the one that minimises the cross-validated prediction error. It additionally included a screening procedure based on running lasso on only the variables selected in a first lasso run.

Table 3 shows that the post-lasso estimator was heavily biased with downwardly biased standard error estimators (given by the default model-based standard errors) as a result of ignoring variable selection uncertainty. The proposal based on post-lasso reduced bias in the estimator, as well its variability, but did not result in a convincing improvement in standard errors. Much better results were found with standard use of the lasso and elastic-net, where the proposal was able to remove bias completely, while also reducing variability further relative to the use of post-lasso. It moreover provided unbiased standard error estimators (which are not available for standard lasso and elastic-net procedures in view of the complex distribution of the estimators they return), leading to nominal coverage of the Wald confidence intervals being attained. The standard lasso and elastic-net estimators were less variable, but this is largely due to shrinkage bias, with coefficients often being set to zero. The use of SuperLearner worsened performance. While less bias was observed with the plug-in estimator, standard errors were very poorly estimated resulting in poor coverage of confidence intervals. Results indicate that larger sample sizes are needed for the proposed estimator based on SuperLearner to perform well in settings with such high-dimensional covariates. In the Supplementary Materials, we provide additional simulation results under complex data-generating mechanisms with misspecified link function, in which we also study the performance of cross-fitting.



**TABLE 3** Simulation results on variable selection: empirical bias (Bias) and standard deviation (Emp SD), sample average of the estimated influence-function based standard errors (Mean SE), and coverage of 95% Wald confidence intervals (Cov) for post-lasso (P-lasso), Lasso, elastic-net and SuperLearner (SL), and the proposed variants thereof (AL)

Est.	n = 200				n = 400			
	Bias	SD	SE	Cov	Bias	SD	SE	Cov
P-lasso	0.15	0.21	0.13	65	0.095	0.14	0.099	73
AL P-lasso	0.072	0.21	0.13	75	0.039	0.13	0.093	83
Lasso	−0.031	0.088			0.010	0.072		
AL Lasso	−0.00011	0.15	0.14	93	−0.00096	0.10	0.099	94
Elastic-net	−0.073	0.051			−0.042	0.041		
AL Elastic-net	−0.0085	0.14	0.14	93	−0.011	0.098	0.096	95
SL	−0.19	0.17	0.019	30	0.0048	0.078	0.00033	0.4
AL SL	0.42	0.22	0.11	13	0.096	0.11	0.096	79

6.4 | Complementary log-log link function

In a final set of simulations, we considered the same main effect estimand (5) with a logit link as in Section 6.1, but now included a complementary log-log link in the data-generating model. The exposure  $A$  and  $L$  were generated as in Section 6.1 and  $Y$  was generated in 4 different ways: (1)  $Y \sim \text{Bern}(1 - \exp(-\exp(0.3A + \delta^T L_{[1:5]})))$  where  $\delta$  is a 5-dimensional unit vector scaled by 1/10; (2)  $Y \sim \text{Bern}(1 - \exp(-\exp(0.3A + \delta^T L_{[1:5]} - 0.025L_1^2)))$ ; (3)  $Y \sim \text{Bern}(1 - \exp(-\exp(0.1L_1A + \delta^T L_{[1:5]})))$ ; and (4)  $Y \sim \text{Bern}(1 - \exp(-\exp(0.025/(1 + \exp(0.1L_3 - 0.1L_2)) + 0.075A/(1 + \exp(-0.1L_2)) + 0.125AL_6 - 0.025L_1^2)))$ .

For each setting, we fitted a generalised linear model with a complementary log-log link function, that was linear (on the complementary log-log scale) in  $A$  and the covariates  $L$ ; the maximum likelihood estimator of the main effect of  $A$  is referred to as ‘MLE-cloglog’. Only in the first setting was this model correctly specified; the maximum likelihood estimator is inconsistent for the parameter  $\beta$  indexing a (correctly specified) partially linear complementary log-log model  $\text{cloglog}\{E(Y|A, L)\} = \beta A + \omega(L)$  in the second setting. In the third and fourth, this estimator converges to a population limit which was approximated via simulation, and which may not be easily interpretable.

We also implemented the same estimator from Section 6.1, developed for the estimand (5) with a logit link. We emphasise that although the link function for generating the data was different to the one used in the considered estimand, nevertheless the estimand remains well defined. For comparison, we also considered the maximum likelihood estimator of the main effect for  $A$  in a logistic regression model that was linear (on the logit scale) in  $A$  and  $L$  (MLE-logit). The logistic model was misspecified in each of the experiments, so bias and coverage of the maximum likelihood estimator were again reported relative to the estimator’s population limit (approximated via simulation). From the results in Table 4, one can see that even when the complementary log-log link was used in the data-generating model, our estimator continues to infer a weighted average of the conditional association of interest (on the log-odds scale) with relatively low bias, and with confidence intervals that possess close to their advertised coverage. This is reassuring, given

**TABLE 4** Simulation results with a misspecified link function: empirical bias (Bias) and standard deviation (SD), sample average of the estimated influence-function based standard errors (SE), and coverage of 95% Wald confidence intervals (Cov). Bias and coverage taken w.r.t. the limiting values of each estimator in experiment 1 (0.3 (MLE-cloglog), 0.51 (MLE-logit), 0.55 (AL)), experiment 3 (0.04 (MLE-cloglog), 0.03 (MLE-logit) and 0.1 (AL)) and experiment 4 (0.29 (MLE-cloglog), 0.42 (MLE-logit) and 0.12 (AL)); in experiment 2, bias/coverage were taken w.r.t. the truth for MLE-cloglog and w.r.t the population limits for the other estimators (0.83 (MLE-logit), 0.52 (AL))

Exp.	Est.	n = 500				n = 1000				n = 2000			
		Bias	SD	SE	Cov	Bias	SD	SE	Cov	Bias	SD	SE	Cov
1	MLE-cloglog	0.01	0.14	0.13	94	0.01	0.10	0.09	93	0.00	0.07	0.07	96
	MLE-logit	0.02	0.25	0.24	94	0.01	0.17	0.17	94	0.00	0.12	0.12	96
	AL	−0.02	0.21	0.20	94	−0.03	0.16	0.15	92	−0.03	0.11	0.11	93
2	MLE-cloglog	0.24	0.14	0.14	58	0.23	0.10	0.10	33	0.23	0.07	0.07	8
	MLE-logit	0.03	0.24	0.23	94	0.01	0.17	0.16	95	0.00	0.11	0.11	95
	AL	0.10	0.21	0.20	90	0.06	0.16	0.14	88	0.03	0.12	0.10	91
3	MLE-cloglog	0.00	0.15	0.14	95	0.00	0.10	0.10	94	0.00	0.07	0.07	95
	MLE-logit	0.00	0.24	0.23	95	0.00	0.16	0.16	95	−0.01	0.11	0.11	95
	AL	0.02	0.21	0.20	94	0.01	0.15	0.14	94	0.01	0.11	0.10	93
4	MLE-cloglog	0.00	0.13	0.14	96	0.01	0.09	0.10	95	0.00	0.07	0.07	95
	MLE-logit	0.01	0.20	0.21	96	0.01	0.14	0.15	95	0.00	0.10	0.10	95
	AL	0.10	0.19	0.20	93	0.06	0.14	0.14	93	0.02	0.10	0.10	94

that often data analysts may prefer to report results on the log-odds (rather than complementary log-log) scale, since the interpretation may be more familiar. While also the results for the other estimators appear favourable because bias is defined relative to their population limit for these estimators, a key drawback of these estimators is that it is not well understood what their population limit represents.

## 7 | DATA ANALYSIS

The First Steps program was set up in 1989 in Washington State, United States, in order to serve low-income pregnant women and children. A specific goal was to reduce the risk of low birth weight. Using data obtained from birth certificates from 2,500 children born in King County, Washington in 2001, we sought to evaluate the effects of the First Steps program on infant birth-weight, as well as its association with maternal age. We were also interested in the possible interaction between the two exposures considered.

We first carried out a more traditional analysis using parametric models. Specifically, we fit a linear model for infant birth weight (in grams), with an indicator of participation on the First Steps program and maternal age as predictors, as well other baseline covariates (child’s sex, mother’s age, race (asian, black, hispanic, white or other), number of previous live born infants, weight prior to pregnancy, education, smoking status and marital status). This model yielded estimates of −13.57 (95% CI: −76.34 to 49.20) for First Steps participation. Assuming that we have adjusted

for all common causes of First Steps participation and birth weight, and additionally that the linear model is correctly specified, then the first regression coefficient suggests that participation in the program led to an average reduction of  $-13.57$  grams in birth weight (although the confidence interval contained the null). For comparison, fitting a linear model unadjusted for covariates yielded an estimate of  $-66.18$  (95% CI:  $-125.79$  to  $-6.57$ ), such that ignoring confounding gives the impression that the intervention was harmful. We then refit the linear model with an interaction term; it was estimated that the association between program participation and birth weight increased by 2.7 units per year increase in maternal age (95% CI:  $-6.99$  to  $12.33$ ). We fit a separate linear model, adjusted for all other covariates except program participation, to assess the effect of age which was estimated as  $0.037$  (95% CI:  $-4.40$  to  $4.47$ ). We did not adjust for participation given that it was an externally introduced factor that may be predicted by age.

We repeated this analysis after dichotomising the outcome (an infant was considered to have low birth weight if they weighed  $<2,500$  g). The estimated log-odds ratios for low birth weight were  $-0.038$  (95% CI:  $-0.55$  to  $0.45$ ) for First Steps participation and  $0.037$  for age (95% CI:  $0.00$ ,  $0.07$ ), again taken from separate models.

We re-analysed the data using the methods proposed in this article; first we estimated the propensity-overlap weighted effect of First Steps participation on birth weight using the influence function-based estimator in Equation (15). The nuisance functionals  $E(A|L)$  and  $E(Y|L)$  (along with all others described in the section) were estimated using the SuperLearner. The SuperLearner library included a generalised linear model with main effects only, a generalised linear model with main effects and pairwise interactions, random forests regression, support vector machines,  $k$ -nearest neighbours and the default generalised additive model procedures as well as 8 additional variants of it with degrees of freedom fixed at 3, ..., 10. We obtained an estimate of  $-5.91$  (95% CI:  $-85.12$  to  $73.31$ ), which was smaller in magnitude than in the previous analysis, and reflects our a priori belief that program participation is unlikely to lead to a strong decrease in infant birth weight. In looking at the weighted effect of maternal age, we again did not adjust for program participation. The proposal yielded an estimate of  $-1.39$  (95% CI:  $-6.32$  to  $3.54$ ). By construction, these can be interpreted as the main effects of First Steps participation and age, regardless of the presence of possible interactions. In a subsequent analysis, we also re-estimated the interaction between First Steps participation and maternal age without making assumptions about possible dependencies between these exposures, and found the interaction to be more pronounced. We obtained an estimate of  $6.96$  (95% CI:  $-6.90$  to  $20.82$ ) based on SuperLearner. Repeating this analysis for the weighted average difference of log-odds of low birthweight gave the effect of program participation as  $0.01$  (95% CI:  $-0.43$  to  $0.44$ ) and maternal age as  $0.055$  (95% CI:  $0.03$  to  $0.08$ ).

## 8 | DISCUSSION

We have emphasised that most data analyses rely on modelling assumptions in more intricate ways than we may realise. They extract information from those assumptions, rather than from the data alone. This may result in estimators for, for instance, a conditional association that are not guaranteed to summarise that association well (e.g. that cannot be viewed as a weighted average of covariate-specific conditional association measures) when those modelling assumptions fail. It may moreover deliver overly optimistic uncertainty assessments, even when based on sandwich standard errors, that are only justified when those modelling assumptions hold. With others, we therefore recommend that the starting point of a data analysis becomes the choice of an estimand,

as opposed to the choice of a model. This ensures that the analysis' aim is unambiguously clear at all times, regardless of issues of model misspecification. It moreover assures that uncertainty assessments, by virtue of being obtained under the nonparametric model, reflect solely the information that is contained in the data. To prevent this rendering interpretation more complicated, we have chosen to focus on estimands that can be interpreted as familiar regression parameters when corresponding models hold, but continue to capture what these parameters aim to summarise when these models are misspecified. The proposal thereby addresses the usual tension between the need for possibly complex models versus the desire to obtain easy-to-communicate results (Breiman, 2001).

The idea of starting the analysis with the choice of an estimand, has become well integrated in causal inference research (Hernan & Robins, 2020). Here, estimands are typically chosen with a view on specific interventions, whose impact one aims to assess. This literature has primarily focused on the average causal effect,  $E(Y^1 - Y^0)$ , which expresses how different the expected outcome would be if all subjects in the population were treated versus untreated. This is useful - in fact, often more useful than the estimands we consider - if such interventions can be conceived. For a continuous exposure, contrasts of  $E(Y^a)$  for different exposure levels  $a$  are arguably less meaningful as interventions that force each one's exposure to take on level  $a$  may not be realistic (consider e.g. the effect of fixing everyone's BMI at 25) and demand enormous extrapolations. Continuous exposures moreover demand a greater need to summarise, which is naturally done by means of so-called marginal structural models in the causal inference literature (Robins et al., 2000), such as

$$E(Y^a) = \alpha + \beta a,$$

for all  $a$ . Weighted least squares regression of  $Y$  on  $A$ , using so-called stabilised weights  $f(A)/f(A|L)$ , then delivers an estimator for  $\beta$  whose probability limit equals

$$\frac{\int \{a - E(A)\} E(Y|A = a, L = l) f(a) f(l) da dl}{\text{Var}(A)}.$$

This expression shows that while the starting point of a causal analysis is often an explicit estimand, also here, the desire to summarise high-dimensional information leads one to working with estimands that are implicitly defined by the estimation procedure. In particular, adjustment for baseline covariates  $C$  is rather common in marginal structural models and has lead one to consider projection estimands for the parameters indexing models like

$$E(Y^a|C) = \alpha + \beta a + \gamma C.$$

These have for instance been defined, for dichotomous exposure, as the minimiser to

$$E[(Y^1 - \alpha - \beta - \gamma C)^2] + E[(Y^0 - \alpha - \gamma C)^2]$$

(Neugebauer & van der Laan, 2007). When stabilised weights are used or a non-linear link function is involved, then this raises similar concerns as in Section 2 when the dependence of  $Y^a$  on  $C$  is misspecified, for then the minimiser for  $\beta$  may no longer be guaranteed to capture the exposure effect on outcome.

In causal inference applications, this explicit need for summarisation can be avoided by focussing on estimands that depend on the natural value of treatment (Hubbard & van der Laan, 2008; Muñoz & van der Laan, 2012; Young et al., 2014), for instance, that consider the effect of shifting the exposure with one unit:

$$E(Y^{A+1} - Y^A).$$

This estimand, which also reduces to  $\beta$  in model (4) with identity link when that model is correctly specified, is directly relevant if interest lies in the effect of interventions that aim to increase the exposure by one unit. In such settings, it is easier to interpret than the estimand (5). It has the drawback, however, that such specific interventions may be rare in practice and that the estimand is very specific to the chosen intervention. In particular, since  $E(Y^{A+2} - Y^A)$  will not generally equal twice  $E(Y^{A+1} - Y^A)$ , a need to summarise the effects  $E(Y^{A+a} - Y^A)/a$  for different levels of  $a$  may remain when there is no convincing reason to consider  $a = 1$ . In this paper, we have therefore opted to work with more generic estimands, that are also relevant when no specific interventions are considered (e.g. when describing the association of an outcome with age, when measuring time trends, ...), and whose efficient influence function under the nonparametric model does not involve inverse weighting by the conditional density of  $A$ , given  $L$ . Such inverse weighting complicates the use of flexible data-adaptive procedures when, for example, the exposure is continuous (it may require the need for binning, as in Muñoz and van der Laan (2012)), as conditional density estimation is a difficult problem which has received little attention in the machine learning literature. Inverse weighting also reflects a change of measure, and thus signals extrapolations being made (e.g. the fact that a one-unit increase in exposure may be very unlikely for subjects in certain covariate strata) and thus estimators that rely on it may exhibit erratic behaviour, even when the density is known. We have therefore focussed on estimands with a generic definition (regardless of whether the exposure is discrete or continuous, and regardless of whether one aims to answer a causal question or not), for which inference can be developed in a generic way (regardless of whether the exposure is discrete or continuous). Such generic estimands are important to enable broadly accessible data analyses. Nevertheless, we acknowledge that in specific circumstances, other estimands may be of greater interest.

Arguably, a drawback of the estimands considered in this paper is that they depend on the exposure distribution, as is for instance seen in Equation (6). This may be considered undesirable (in a similar way that the partial likelihood estimator of the hazard ratio under a Cox model has been criticised for its limit depending on the censoring distribution in a complicated manner (van der Laan & Rose, 2011); however, it is the unavoidable consequence of working with estimands that eschew inverse weighting and thus avoid strong extrapolations away from the observed exposure distribution.

In our attempt to come up with generic estimands for regression parameters, we have experienced a need for clear principles for choosing estimands, as opposed to letting them be mere projection parameters (Buja et al., 2019b). In the considered context, we have found it useful to start from the premise that  $E(Y|A, L)$  is known for all levels of  $A$  and  $L$ , and to consider how to best summarise this information when the aim is parsimony. This is best done with some regression model in mind, to ensure that the estimand coincides with a familiar regression parameter when that model is correctly specified, and thus remains well interpretable. To prevent that the assumptions embodied in the entire regression model dominate the choice of estimand, we have focussed on (generalised) partially linear models, which merely specify the conditional

association or effect modification term of interest. The population limit of semiparametric estimators under such model may then serve as a template for a choice of estimand. Such choice is non-unique. In our work, we have aimed for simplicity, realising that other estimands (e.g. that involve inverse weighting by the conditional outcome variance) can be inferred more efficiently. For instance, when  $g(\cdot)$  is the logit link, it may be advantageous to define the main effect estimand instead as

$$\frac{E\left(\sigma^2(A, L) \left[A - \frac{E\{\sigma^2(A, L)A|L\}}{E\{\sigma^2(A, L)|L\}}\right] g\{E(Y|A, L)\}\right)}{E\left(\sigma^2(A, L) \left[A - \frac{E\{\sigma^2(A, L)A|L\}}{E\{\sigma^2(A, L)|L\}}\right] A\right)},$$

for  $\sigma^2(A, L) = E(Y|A, L)\{1 - E(Y|A, L)\}$ . Further work is needed to develop inference for this estimand, and insight in its interpretation.

In future work, we will make similar developments for parameters indexing proportional hazard models for time-to-event data and marginal models for repeated measures data. We will moreover study how the dependence of  $E(Y|A = a, L = l)$  for continuous  $a$  can be described in a less restrictive way by focussing on the estimand

$$\frac{E[\text{Var}(A|L)g\{E(Y|A = a, L)\}]}{E\{\text{Var}(A|L)\}}$$


as an unrestricted function of  $a$ .

## ACKNOWLEDGEMENTS

The authors thank Mark van der Laan for inspiring discussions that have influenced this work, and Vanessa Didelez, Oliver Hines and Pawel Morzywolek for useful feedback. The authors also thank the support from BOF Grants BOF.01P08419 and BOF.24Y.2017.0004.01.

## ORCID

Stijn Vansteelandt  <https://orcid.org/0000-0002-4207-8733>

Oliver Dukes  <https://orcid.org/0000-0002-9145-3325>

## REFERENCES

- Angrist, J.D. & Krueger, A.B. (1999) Empirical strategies in labor economics. In: Ashenfelter, O.C. & Card, D. (Eds.) *Handbook of labor economics*, vol. 3, Amsterdam: Elsevier, pp. 1277–1366.
- Angrist, J.D. & Pischke, J.-S. (2009) *Mostly harmless econometrics: an empiricist's companion*. Princeton: Princeton University Press. OCLC: ocn231586808.
- Aronow, P.M. & Samii, C. (2016) Does regression produce representative estimates of causal effects? *American Journal of Political Science*, 60, 250–267.
- Athey, S., Tibshirani, J. & Wager, S. (2019) Generalized random forests. *The Annals of Statistics*, 47, 1148–1178.
- Belloni, A., Chernozhukov, V. & Wei, Y. (2013) Honest confidence regions for a regression parameter in logistic regression with a large number of controls. *Tech. rep.*, cemmap working paper.
- Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013) Valid post-selection inference. *The Annals of Statistics*, 41, 802–837.
- Bickel, P.J., Klaassen, C.A., Bickel, P.J., Ritov, Y., Klaassen, J., Wellner, J.A. et al. (1993) *Efficient and adaptive estimation for semiparametric models*, vol. 4. Baltimore: Johns Hopkins University Press Baltimore.
- Breiman, L. (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16, 199–231.

- Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M. et al. (2019a) Models as approximations I: consequences illustrated with linear regression. *Statistical Science*, 34, 523–544.
- Buja, A., Brown, L., Kuchibhotla, A.K., Berk, R., George, E., Zhao, L. et al. (2019b) Models as approximations II: a model-free theory of parametric regression. *Statistical Science*, 34, 545–565.
- Buja, A., Kuchibhotla, A.K., Berk, R., George, E., Tchetgen Tchetgen, E. & Zhao, L. (2019c) Models as approximations—rejoinder. *Statistical Science*, 34, 606–620.
- Chambaz, A., Neuvial, P. & van der Laan, M.J. (2012) Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6, 1059–1099.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duo, E., Hansen, C., Newey, W. et al. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21, C1–C68.
- Crump, R.K., Hotz, V.J., Imbens, G.W. & Mitnik, O.A. (2006) Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. *Tech. rep.*, National Bureau of Economic Research.
- Freedman, D.A. (2006) On the so-called “Huber Sandwich Estimator” and “Robust Standard Errors”. *The American Statistician*, 60, 299–302.
- Graham, B.S. & Pinto, C.C.D.X. (2018) Semiparametrically efficient estimation of the average linear regression function. *arXiv:1810.12511 [econ]*. Available from: <http://arxiv.org/abs/1810.12511>. ArXiv: 1810.12511.
- Hernan, M.A. & Robins, J.M. (2020) *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.
- Hubbard, A.E. & van der Laan, M.J. (2008) Population intervention models in causal inference. *Biometrika*, 95, 35–47.
- Kennedy, E.H., Lorch, S. & Small, D.S. (2019) Robust causal inference with continuous instruments using the local instrumental variable curve. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81, 121–143.
- Kosorok, M.R. (2007) *Introduction to empirical processes and semiparametric inference*. Berlin: Springer Science & Business Media.
- van der Laan, M.J. & Rose, S. (2011) *Targeted learning*. Springer Series in Statistics. New York, NY: Springer New York.
- van der Laan, M.J. & Rubin, D. (2006) Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2.
- van der Laan, M.J., Polley, E.C. & Hubbard, A.E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6.
- Lin, W. (2013) Agnostic notes on regression adjustments to experimental data: reexamining Freedman’s critique. *The Annals of Applied Statistics*, 7, 295–318.
- Muñoz, I.D. & van der Laan, M. (2012) Population intervention causal effects based on stochastic interventions. *Biometrics*, 68, 541–549.
- Neugebauer, R. & van der Laan, M. (2007) Nonparametric causal effects based on marginal structural models. *Journal of Statistical Planning and Inference*, 137, 419–434.
- Newey, W.K. & Robins, J.R. (2018) Cross-Fitting and Fast Remainder Rates for Semiparametric Estimation. *arXiv:1801.09138 [math, stat]*. Available from: <http://arxiv.org/abs/1801.09138>. ArXiv: 1801.09138.
- Nie, X. & Wager, S. (2017) Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*.
- Pfanzagl, J. (1990) Estimation in semiparametric models. In: *Estimation in semiparametric models*. Berlin: Springer, pp. 17–22.
- Robins, J.M., Mark, S.D. & Newey, W.K. (1992) Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 479–495.
- Robins, J.M., Hernan, M.A. & Brumback, B. (2000) Marginal structural models and causal inference in epidemiology.
- Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. (2008) Higher order influence functions and minimax estimation of nonlinear functionals. In: *Probability and statistics: essays in honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics.
- Robinson, P.M. (1988) Root-n-consistent semiparametric regression. *Econometrica: Journal of the Econometric Society*, 931–954.



- Rotnitzky, A. & Robins, J. (1997) Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine*, 16, 81–102.
- Rotnitzky, A., Robins, J.M. & Scharfstein, D.O. (1998) Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93, 1321–1339.
- Słoczyński, T. (2020) Interpreting OLS estimands when treatment effects are heterogeneous: smaller groups get larger weights. *The Review of Economics and Statistics*, 1–27. Available from: [https://direct.mit.edu/rest/article/doi/10.1162/rest\\_a\\_00953/97692/Interpreting-OLS-Estimands-When-Treatment-Effects](https://direct.mit.edu/rest/article/doi/10.1162/rest_a_00953/97692/Interpreting-OLS-Estimands-When-Treatment-Effects)
- Scharfstein, D.O., Rotnitzky, A. & Robins, J.M. (1999) Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94, 1096–1120.
- Tan, Z. (2019) On doubly robust estimation for logistic partially linear models. *Statistics and Probability Letters*, 155, 108577.
- Tchetgen Tchetgen, E.J. (2013) On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. *American Journal of Epidemiology*, 177, 1314–1316.
- Tchetgen Tchetgen, E.J., Robins, J.M. & Rotnitzky, A. (2010) On doubly robust estimation in a semiparametric odds ratio model. *Biometrika*, 97, 171–180.
- Vansteelandt, S. & Daniel, R.M. (2014) On regression adjustment for the propensity score. *Statistics in Medicine*, 33, 4053–4072.
- Vansteelandt, S. & Joffe, M. (2014) Structural nested models and g-estimation: the partially realized promise. *Statistical Science*, 29, 707–731.
- Vansteelandt, S., VanderWeele, T.J., Tchetgen, E.J. & Robins, J.M. (2008) Multiply robust inference for statistical interactions. *Journal of the American Statistical Association*, 103, 1693–1704.
- Wasserman, L. (2014) Discussion: “A significance test for the lasso”. *The Annals of Statistics*, 42, 501–508.
- White, H. (1980) A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48, 817.
- Whitney, D., Shojaie, A. & Carone, M. (2019) Comment: models as (deliberate) approximations. *Statistical Science*, 34, 591–598.
- Young, J.G., Hernán, M.A. & Robins, J.M. (2014) Identification, estimation and approximation of risk under interventions that depend on the natural value of treatment using observational data. *Epidemiologic Methods*, 3, 1–19.
- Zheng, W. & van der Laan, M.J. (2011) Cross-validated targeted minimum-loss-based estimation. In: van der Laan, M.J. & Rose, S. (Eds.) *Targeted learning*. New York, NY: Springer New York, pp. 459–474.

## SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

**How to cite this article:** Vansteelandt, S. & Dukes, O. (2022) Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84(3), 657–739. Available from: <https://doi.org/10.1111/rssb.12504>

## DISCUSSION CONTRIBUTIONS

# Proposer of the vote of thanks to Vansteelandt and Dukes and contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’

Rhian M. Daniel

Division of Population Medicine, Cardiff University, Cardiff, UK

## Correspondence

Rhian M. Daniel, Division of Population Medicine, Cardiff University, Cardiff, UK.

Email: [danielr8@cardiff.ac.uk](mailto:danielr8@cardiff.ac.uk)

## 1 | TWO CONTRASTING PHILOSOPHIES

Traditional statistical modelling starts from a family  $\mathcal{F}$  of observed data laws indexed by unknown parameters of interest  $\beta$ . The goal is to make inference about  $\beta$  under the assumption that  $\mathcal{F}$  contains the true law. By labelling  $\beta$  ‘of interest’, it is implied that  $\mathcal{F}$  can be expressed such that  $\beta$  naturally encompasses the main scientific goal, which is not always the case. Furthermore (e.g. ch.1 Cox & Hinkley, 1979) if  $\mathcal{F}$  does not contain the truth then the inferential theory loses relevance and the interpretation of  $\beta$  is obscure. Such concerns sensibly lead to model checking procedures, which themselves raise further concerns, as VD describe.

The *causal inference* and *targeted learning* schools (Hernán & Robins, 2020; van der Laan & Rose, 2011) start instead from an estimand, chosen to reflect the scientific question, without reference to any statistical model. Subsequent estimation and inference are tailored to this estimand, sometimes using a parametric model  $\mathcal{F}$ , but not to define the estimand. The targeted learning framework advocates replacing  $\mathcal{F}$  with machine learning algorithms, using the estimand’s influence function and accompanying theory to derive estimators with well-understood asymptotic behaviour.

Vansteelandt and Dukes (henceforth VD) propose a practical resolution to an important tension between two philosophies of statistical inference. I summarise these aspects before discussing how we might revise our understanding of ‘bias–variance trade-off’ in statistical modelling in the light of VD’s work.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Author. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

Although the hygiene of the latter approach is eminently attractive, its implementation requires statistical expertise. In principle, each bespoke estimand demands that all subsequent steps be derived afresh, with no guarantees that the resulting estimator has good properties (e.g. when the estimand is too ambitious given the available data). Practical applications of targeted learning thus tend to focus on simple estimands (e.g. the marginal effect of a binary exposure) where off-the-shelf implementations are readily available. This leaves users in a quandary when their scientific question is more complex, for example when the exposure is continuous, as in the settings considered by VD.

## 2 | THE BEST OF BOTH WORLDS

VD start, as in the traditional approach, from a generalised linear model  $\mathcal{F}$  indexed by  $\beta$ . This has the advantage of restricting attention to quantities that are plausibly reliably estimable from the data. For any estimator  $\hat{\beta}$ , consistent under  $\mathcal{F}$ , their philosophy is to consider its probability limit  $\beta^*$  under  $\mathcal{A}$ , the set of *all* possible data laws. The honest estimand  $\beta^*$  is only considered acceptable if it corresponds (under  $\mathcal{A}$ ) to a weighted average of parameters  $\beta_l$ , where each  $\beta_l$  has the interpretation of  $\beta$  restricted to levels  $l$  of some covariates  $L$ , not of primary interest.

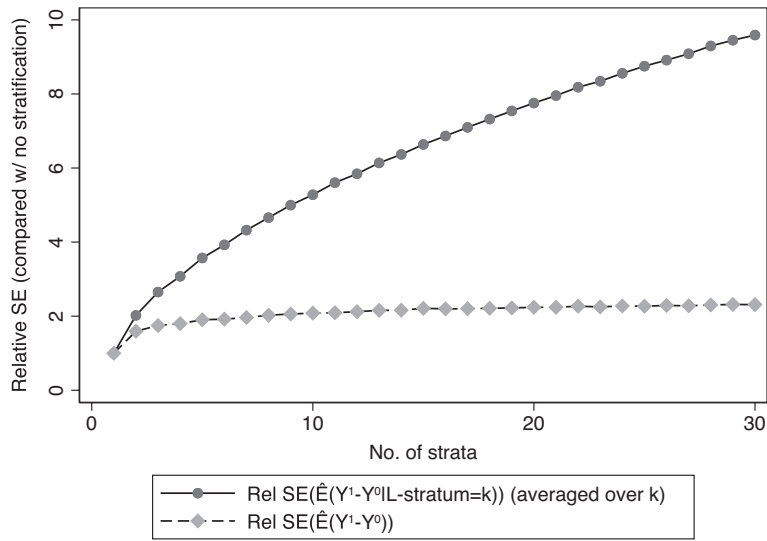
VD set high standards for making inference about  $\beta^*$ , namely consistent estimation and parametric convergence rates under ‘lean’ regularity assumptions, honest inference after algorithm/variable selection, and no density estimation for continuous variables. They argue convincingly that such demands are necessary for the data to speak for themselves about  $\beta^*$ , and describe a general procedure that meets these standards in the case of any parameter motivated by a GLM.

## 3 | PRECISION IS BOUGHT WITH BLUNTNESS NOT BIAS

That VD propose essentially non-parametric estimation may seem alarming in the light of the curse of dimensionality (Stone, 1985). Indeed, the traditional approach based on one simple model is often justified on the grounds of a bias–variance trade-off: we assume a simple (‘wrong but useful’) model since it buys precision in modest-sized data sets. The simulation studies presented by VD illustrate that this intuition is faulty: their assumption-lean estimators are also relatively precise, but then at what cost?

Our intuition was developed in the context of the traditional approach in which  $\mathcal{F}$  plays the two roles described by VD: (i) estimand definition, and (ii) representing the set of possible data laws. Traditionally, choosing a more complex model  $\mathcal{F}$  leads simultaneously to a less parsimonious estimand and a larger set of possible data laws. VD, on the other hand, propose a parsimonious estimand, coupled with only very lean restrictions on the set of data laws: parsimony in the first sense but not the second. Figure 1 gives a simple illustration of how parsimony in both senses increases efficiency, but with parsimony of type (i) having a greater impact than type (ii).

Since consistent estimation is guaranteed under very lean assumptions, and thus bias essentially avoided, the sacrifice made by VD’s parsimony (in the first sense) with which they buy precision is, I believe, not bias but bluntness. A more nuanced (less blunt) understanding of, say, a continuous exposure’s effect on an outcome, could be gained by choosing a less parsimonious summary, for example one that separately summarises the effect in more sub-groups, but at the cost of increased variance.



**FIGURE 1** This graph shows the increase in relative standard error for the estimator of two different types of estimands after subdividing a covariate  $L$  into progressively more strata. The 3000 simulated datasets each with sample size 1000 are from a hypothetical observational study with a continuous confounder  $L \sim N(0, 1)$ , a binary exposure  $A$  with  $\Pr(A = 1 | L) = \text{expit}(L)$  and a binary outcome  $Y$  with  $\Pr(Y = 1|A, L) = \text{expit}(-2 + 0.2AL^2)$ . Each dataset is divided into an increasing number  $s$  of approximately equally-populated strata based on the observed quantiles of  $L$ . We first plot the empirical standard deviation of the stratum-specific estimator of the average causal effect in each stratum separately, when splitting into  $s = 1, \dots, 30$  strata relative to 1 (i.e. no stratification). Since the SE varies by stratum, the plot in fact takes the average of the SE over the  $s$  strata. We then plot the relative empirical standard deviation of the estimator of the average causal effect (marginalised over the strata) when the data analysis model is stratified into  $s$  strata relative to 1. Since the true model for  $Y$  given  $A$  and  $L$  has the same form regardless of the value of  $L$ , the models with and without stratification are all correctly specified. This allows us to explore, on the one hand, the impact of needless flexibility in  $\mathcal{F}$  in the sense described in (ii) in the text (the slowly increasing lower line) compared with the additional impact of decreasing parsimony in the estimand of interest, i.e. sense (i) in the text (the more steeply increasing upper line)

#### 4 | CONCLUDING REMARKS

VD start from the viewpoint that the two approaches in Section 1 are unsatisfactory. The traditional well-trodden path offers a comfortable ride but often to an unknown and uninteresting destination with a dishonest account of how we got there. On the other hand, the targeted learning path, in aiming admirably for the summit of a yet-to-be-conquered mountain, is often too perilous to navigate with our modest equipment and abilities. VD offer a third way, which feels on the surface much like the first, but leads to a well-defined destination that is both practically reachable and at least somewhere in the foothills of scientific interest. Beneath the surface lies much of the sophisticated technology from the targeted learning journey, but as passengers we need not necessarily know how to operate it, thanks to their general-purpose solution.

I conclude by congratulating Vansteelandt and Dukes on their innovative yet pragmatic proposal presented in a wonderfully didactic fashion that provokes us to rethink fundamental aspects of statistical modelling. I enthusiastically propose the vote of thanks.

## REFERENCES

- Cox, D.R. & Hinkley, D.V. (1979) *Theoretical statistics*. Boca Raton: CRC Press.
- Hernán, M.A. & Robins, J.M. (2020) *Causal inference: what if*. Boca Raton: Chapman & Hall/CRC.
- Stone, C.J. (1985) Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2), 689–705.
- van der Laan, M.J. & Rose, S. (2011) *Targeted learning*. Berlin: Springer.

**How to cite this article:** Daniel, R.M. (2022) Proposer of the vote of thanks to Vansteelandt and Dukes and contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 686–689. Available from: <https://doi.org/10.1111/rssb.12513>

DOI: 10.1111/rssb.12514

# Seconder of the vote of thanks to Vansteelandt and Dukes and contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’

Vanessa Didelez<sup>1,2</sup>

<sup>1</sup>Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany

<sup>2</sup>Department of Mathematics and Computer Science, University of Bremen, Bremen, Germany

## Correspondence

Vanessa Didelez, Leibniz Institute for Prevention Research and Epidemiology - BIPS, Bremen, Germany.

Email: [vdidelez@uni-bremen.de](mailto:vdidelez@uni-bremen.de)

In my view, one of the most important contributions of the field of causal inference has been to place the target of inference, the desired estimand, at the centre of the analysis. The estimand is chosen in view of the research question, and typically reflects what decision problem we need to solve or what our ideal (target) trial would be. Crucially, the (causal) estimand is *not* automatically a parameter that happens to parametrise a chosen model. The role of models is mostly as

---

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Author. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

a mere tool. For instance, with a survival outcome we may employ hazard regressions but still obtain effects in terms of survival probabilities.

This is in some contrast to traditional ‘statistical modelling’, where the model often appears to be the starting point, somehow suggested by the data (e.g. a logistic regression for a binary outcome), and the natural coefficients are reported (e.g. log-odds-ratios). However, it has become abundantly clear that, as far as *prediction* is concerned, default regression models are regularly outperformed by machine learning methods such as random forests and similar flexible methods in competitions. The former are then typically defended with the argument of being more interpretable and their parameters being useful summaries of (conditional) dependence.

The thoughtful proposal of Vansteelandt and Dukes (V&D) addresses many of these issues with an important lesson right at the start: models, even when only used as tools, may implicitly affect the meaning of our estimands and the desired summaries may be invalidated under misspecification. V&D therefore place a (particular) estimand at the core, aiming at a simple summary of a (high-dimensional) conditional dependence such that the estimand remains sensible regardless of whether a specific model holds. The influence function-based estimation method then uses flexible machine learning for the nuisance functions in a way that ensures valid inference.

While I agree with many of V&D’s points, I am (very slightly) concerned that they might distract from asking scientifically relevant questions, which would conflict with the authors’ intention. Their proposal restricts our choice of estimand. For example, when the exposure  $A$  is continuous, V&D make the entirely valid point that the potential outcome  $Y^a$  (and thus an estimand as  $E(Y^a - Y^{a'})$ ) typically represents an unrealistic intervention of setting  $A$  to exactly  $a$  even for people whose ‘natural’ value of  $A$  would be very different from  $a$ . But, such an estimand can be scientifically meaningful, for example when  $A$  is the dosage of a drug; in contrast, it is less meaningful when  $A$  is BMI or income, for example. However, the proposed alternative estimand does not solve this problem—only when we actually *formulate* a scientifically relevant question will we (possibly) find scientifically relevant answers. Giving less weight to problematic covariate regions does not achieve this—and it does not absolve us from trying harder to elicit what a scientifically relevant estimand might be instead, for instance for the effect of BMI on an outcome. Moreover, it is a useful feature of approaches like IPW or propensity score matching that they alert us to problems, for example when we carry out diagnostics and find that there is a lack of overlap and then go on to characterise regions of sufficient overlap; or even when we just have extreme weights and confidence intervals get very wide, this rightly indicates that no useful statement can be made about our estimand because there is too little information in the data. I worry that such aspects are perhaps lost with V&D’s approach, or it would be interesting to know if it could be supplemented by something analogous.

In motivating the proposed estimands, V&D further refer to our ‘familiarity’ with main effects and interactions in GLMs. However, familiarity is not per-se a relevant criterion: it does, again, not ensure scientific relevance. There are plenty of examples, for instance the problematic causal interpretation of hazard ratios despite many medical statisticians being extremely familiar with them. Besides, what exactly is it that we think we are familiar with? When regression models are used to describe conditional dependencies, the correct interpretation of regression coefficients as just a measure of conditional dependence, and not as an effect, is rare. A clear distinction between covariates included to adjust for confounding and those included as potential ‘effect’ modifiers is also rare, typically no rationale for the inclusion of particular covariates is given at all—what then motivates the choice of  $L$  in the basic quantity  $E(Y|A, L)$ ?

In summary, I very much welcome V&D’s proposal as a huge improvement on traditional statistical regression modelling; but with regard to causal analyses there is still room for further

research into combining the truly impressive results on assumption-lean valid inference with the quest for more scientifically meaningful estimands. No doubt, V&D will lead the way.

It is with greatest pleasure that I second the vote of thanks for this most stimulating and important paper.

**How to cite this article:** Didelez, V. (2022) Second of the vote of thanks to Vansteelandt and Dukes and contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 689–691. Available from: <https://doi.org/10.1111/rssb.12514>

The vote of thanks was passed by acclamation.

DOI: 10.1111/rssb.12515

## Peng Ding’s contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’ by Vansteelandt and Dukes

### Peng Ding

Department of Statistics, University of California Berkeley, Berkeley, California, USA

#### Correspondence

Peng Ding, Department of Statistics, University of California Berkeley, Berkeley, CA, USA.

Email: [pengdingpku@berkeley.edu](mailto:pengdingpku@berkeley.edu)

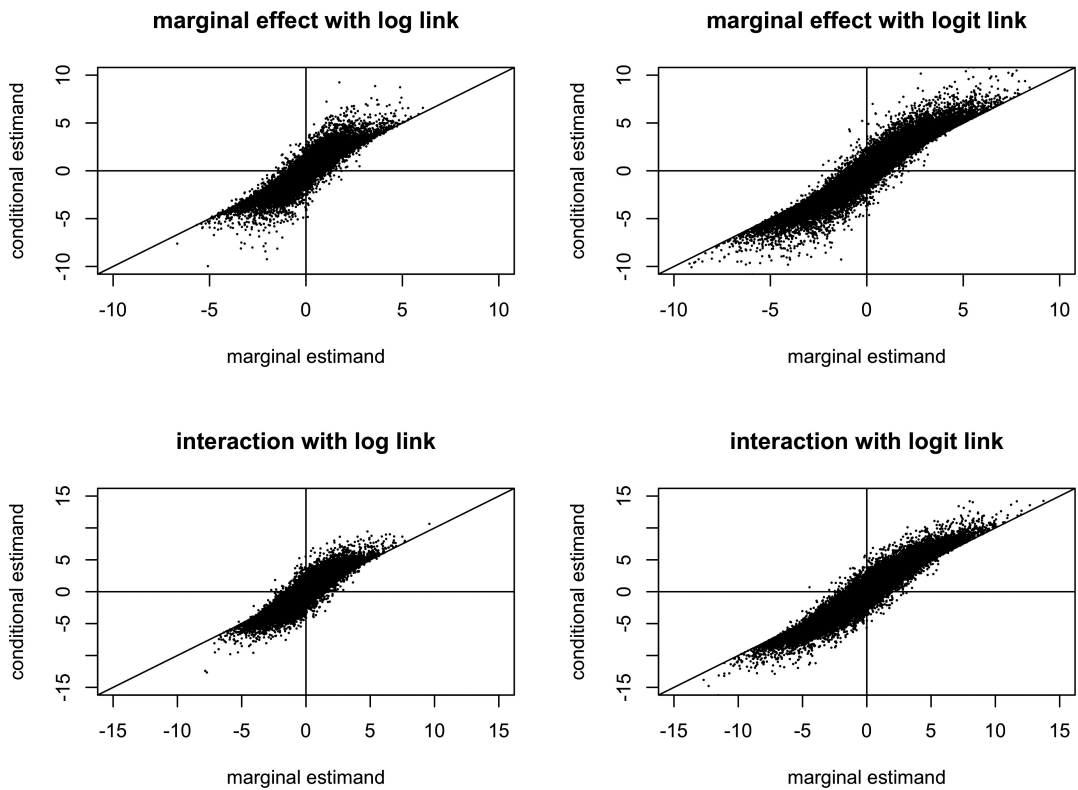
I want to congratulate the authors on a thought-provoking paper on statistical inference with misspecified models. It deepens our understanding of the role of models in causal inference. I offer two comments.

First, the conditional estimands proposed in this paper can be misleading even in randomised experiments. With a binary exposure  $A$ , the authors propose the conditional estimand

$$\tau_{\text{conditional}} = E \left( w_o(L) \left[ g\{E(Y|A = 1, L)\} - g\{E(Y|A = 0, L)\} \right] \right),$$

where  $w_o(L) = \pi(L)\{1 - \pi(L)\}/E[\pi(L)\{1 - \pi(L)\}]$  is the weighting function depending on the propensity score  $\pi(L) = P(A = 1|L)$ . With the identify link  $g(z) = z$ , it reduces to the average treatment effect for the population under the overlap weight (Li et al., 2018). This form of weighting appeared in different settings (e.g. Angrist, 1998; Crump et al., 2006; Vansteelandt & Daniel, 2014; Wallace &





**FIGURE 1** Comparison of the conditional and marginal estimands based on  $10^5$  Monte Carlo samples. The parameters of the distributions of  $L$  and  $Y$  are all drawn independently from  $\text{Uniform}(0, 1)$

Moodie, 2015; Ding, 2021). With non-linear log or logit links for a binary  $Y$ , it reduces to a weighted average of the conditional risk ratio or odds ratio, which does not equal to the standard marginal estimand

$$\tau_{\text{marginal}} = g \left[ E \{ w_0(L) E(Y|A = 1, L) \} \right] - g \left[ E \{ w_0(L) E(Y|A = 0, L) \} \right].$$

The authors point out this well-known issue in their paper. Figure 1 compares  $\tau_{\text{conditional}}$  and  $\tau_{\text{marginal}}$  as well as their analogues for interaction under a completely randomised experiment with a binary  $L$ . In general,  $\tau_{\text{conditional}} \neq \tau_{\text{marginal}}$  and they may not have the same sign. The authors argue that an advantage of  $\tau_{\text{conditional}}$  is that it reduces to  $\beta$  under model (4). In contrast,  $\tau_{\text{marginal}} \neq \beta$  under model (4) unless  $g(z) = z$  or  $g(z) = \log z$ . However,  $\beta$  may not be of interest even if model (4) is correctly specified and even if the data arise from a randomised experiment. It measures the conditional risk ratio or odds ratio given  $L$ , but the policy-relevant estimand is often the marginal risk difference, risk ratio or odds ratio. Freedman (2008) gave a critical assessment of logistic regression in randomised experiments.

Second, with two binary exposures  $A_1$  and  $A_2$ , we may be interested in the main effects and interaction simultaneously. Although model (7) is a convincing motivation for the novel interaction measure, it is not for the main effects measures. Let  $\mathcal{A} = \{(1, 1), (1, 0), (0, 1), (0, 0)\}$ . Section 4 motivates to define the general weighted contrast

$$\tau_c = E \left[ w_0(L) \sum_{(a_1, a_2) \in \mathcal{A}} c(a_1, a_2) g\{E(Y|A_1 = a_1, A_2 = a_2, L)\} \right],$$

where  $w_0(L)$  is the normalised

$$\left\{ \sum_{(a_1, a_2) \in \mathcal{A}} \frac{1}{\pi_{(a_1, a_2)}(L)} \right\}^{-1},$$

also called the general overlap weight (Li & Li, 2019). With  $c(1, 1) = 1$ ,  $c(1, 0) = -1$ ,  $c(0, 1) = -1$ ,  $c(0, 0) = 1$ , it reduces to the interaction measure proposed in Section 4; with  $c(1, 1) = 1/2$ ,  $c(1, 0) = 1/2$ ,  $c(0, 1) = -1/2$ ,  $c(0, 0) = -1/2$ , it reduces to a measure of the main effect of  $A_1$ ; with  $c(1, 1) = 1/2$ ,  $c(1, 0) = -1/2$ ,  $c(0, 1) = 1/2$ ,  $c(0, 0) = -1/2$ , it reduces to a measure of the main effect of  $A_2$ . Following Zhao and Ding (2022), it is straightforward to extend the definition of  $\tau_c$  to the setting with multiple factors  $A_1, \dots, A_K$  ( $K \geq 2$ ). It is a curious question to find the corresponding working model for a contrast  $\tau_{c^*}$ , and more interestingly, for multiple contrasts  $\tau_{c_m}$  ( $m = 1, \dots, M$ ) simultaneously.

The authors provide inspiring derivations of the novel average effect and interaction measures under possibly misspecified models. I am wondering how general their strategy is beyond these two measures.

## REFERENCES

- Angrist, J.D. (1998) Estimating the labor market impact of voluntary military service using social security data on military applicants. *Econometrica*, 66, 249–288.
- Crump, R., Hotz, V.J., Imbens, G. & Mitnik, O. (2006) Moving the goalposts: Addressing limited overlap in the estimation of average treatment effects by changing the estimand. Technical report, National Bureau of Economic Research.
- Ding, P. (2021) The Frisch-Waugh-Lovell theorem for standard errors. *Statistics & Probability Letters*, 168, 108945.
- Freedman, D.A. (2008) Randomization does not justify logistic regression. *Statistical Science*, 23, 237–249.
- Li, F. & Li, F. (2019) Propensity score weighting for causal inference with multiple treatments. *Annals of Applied Statistics*, 13, 2389–2415.
- Li, F., Morgan, K.L. & Zaslavsky A.M. (2018) Balancing covariates via propensity score weighting. *Journal of the American Statistical Association*, 113, 390–400.
- Vansteelandt S. & Daniel, R.M. (2014) On regression adjustment for the propensity score. *Statistics in Medicine*, 33, 4053–4072.
- Wallace, M.P. & Moodie, E.E.M. (2015) Doubly-robust dynamic treatment regimen estimation via weighted least squares. *Biometrics*, 71, 636–644.
- Zhao, A. & Ding, P. (2022) Regression-based causal inference with factorial experiments: estimands, model specifications, and design-based properties. *Biometrika*, in press.

**How to cite this article:** Ding, P. (2022) Peng Ding's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 691–693. Available from: <https://doi.org/10.1111/rssb.12515>

DOI: 10.1111/rssb.12516

# Mats J. Stensrud and Aaron L. Sarvet's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Mats J. Stensrud | Aaron L. Sarvet

École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

## Correspondence

Mats J. Stensrud, École Polytechnique Fédérale de Lausanne, Switzerland.

Email: [mats.stensrud@epfl.ch](mailto:mats.stensrud@epfl.ch)

We congratulate Vansteelandt and Dukes (V & D) with their innovative and interesting article. Here we further explore the interpretation of V & D's main effects estimand.

**An algorithm for causal inference.** There has been tremendous progress in causal inference by approaching causal queries in the following way (Hernan & Robins, 2020; Richardson & Robins, 2013; Robins, 1986):

1. Choose a causal target: an estimand that corresponds to a scientific question of interest (usually a causal parameter).
2. Specify a set of (reasonable) assumptions to define a model, and deduce a functional that equals the causal target at every law in the model.
  - a. If the functional is fully comprised of observed data parameters, then we have full identification.
  - b. If the functional is partially comprised of observed data parameters, then we have partial identification.
3. Implement an estimator of the observed components of the identification functional.

An analysis that results in partial identification, that is, it goes through 2.b, might be considered the classical 'assumption-lean' approach. The assumptions and the causal target are fixed. The investigators leverage their (limited) assumptions to derive bounds, that is 'worst-cases', for the value of the causal target.

---

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](#) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

V & D do not follow this algorithm. Their algorithm fixes a functional of observed data—what they call an ‘estimand’. Then they consider an assumption (a parametric modelling assumption) that may or may not hold, and they derive what we might call a lower bound on the interpretation of their ‘estimand’.

For example, V & D’s Expression (5) corresponds to an average treatment effect under usual identification assumptions *and* a strong parametric assumption. In the absence of the strong parametric assumption, it corresponds to a variance weighted treatment effect (Robins et al., 2008).

But, those who make decisions based on data cannot avoid discrete commitments about assumptions in their analyses. V & D’s algorithm does provide novel information (in the form of a worst-case interpretation) when the investigator is not willing to commit to the (parametric) model assumption. However, in this case decision-makers must be aware of the implications of the worst-case interpretation and how it differs from a causal target that more naturally corresponds to their scientific question of interest. Failure to do so could lead to wrong conclusions, as illustrated in the following example.

**A simple illustration.** Following V & D, consider a binary treatment  $A \in \{0, 1\}$ , a binary covariate  $L \in \{0, 1\}$  and an outcome  $Y \in \mathbb{R}$ . Let  $P(L = 1) = 0.5$ ,  $\pi(1) = p$  and  $\pi(0) = q$ . Suppose that  $Y$  is determined by the structural equation

$$Y = \alpha + \beta(-1)^L A,$$

For  $\alpha \in \mathbb{R}$  and  $\beta > 0$ , the average treatment effect (ATE) is

$$\mathbb{E}(Y^{a=1}) - \mathbb{E}(Y^{a=0}) = \beta \{P(L = 0) - P(L = 1)\} = 0,$$

regardless of  $p$  and  $q$ . However, V & D’s main effects estimand is

$$VD(p, q) = \frac{\mathbb{E}[\pi(L) \{1 - \pi(L)\} (Y^{a=1} - Y^{a=0})]}{\mathbb{E}[\pi(L) \{1 - \pi(L)\}]} = \frac{\{q(1 - q) - p(1 - p)\} \beta}{q(1 - q) + p(1 - p)}.$$

Thus

- $VD(p, q) > 0$  when  $|q - 0.5| < |p - 0.5|$ , and
- $VD(p, q) < 0$  when  $|q - 0.5| > |p - 0.5|$ .

This example illustrates a more general point. V & D’s main effects estimand does not necessarily correspond to an ATE, even when the ATE is 0 and even when there is no unmeasured confounding. In more complex settings, with higher dimensional treatments and non-linear link functions, these issues can still exist.

## REFERENCES

- Hernan, M.A. & Robins, J.M. (2020) *Causal inference: what if*. Boca Raton: Chapman & Hall CRC.
- Richardson, T.S. & Robins, J.M. (2013) Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper, 128(30), 2013.
- Robins, J. (1986) A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9–12), 1393–1512.

Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. (2008) Higher order influence functions and minimax estimation of nonlinear functionals. In: Nolan, D. & Speed, T. (Eds.) *Probability and statistics: essays in honor of David A. Freedman*. Beachwood: Institute of Mathematical Statistics, pp. 335–421.

**How to cite this article:** Stensrud, M.J. & Sarvet, A.L. (2022) Mats J Stensrud and Aaron L. Sarvet's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 694–696. Available from: <https://doi.org/10.1111/rssb.12516>

DOI: 10.1111/rssb.12517

# Heather Battey's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

## Heather Battey

Department of Mathematics, Imperial College London, London, UK

### Correspondence

Heather Battey, Department of Mathematics, Imperial College London, London, UK.

Email: [h.battey@imperial.ac.uk](mailto:h.battey@imperial.ac.uk)

For situations in which there is uncertainty over the underlying probabilistic model, there are at least three broad approaches. One is to seek reliable inference for interest parameters or perhaps, as the authors advocate, for quantities retaining at least a degree of interpretability under misspecification. Another is to acknowledge more explicitly the model uncertainty. A third approach, loosely connected to the first, is to encapsulate uncertainty over the model in a possibly large number of nuisance parameters, to be eliminated in the analysis by suitable conditioning arguments or other problem-specific manoeuvres (e.g. Bartlett, 1937). A helpful example is the use of partial likelihood to evade the baseline hazard function (an infinite-dimensional nuisance parameter) of a proportional hazards model. The appropriateness of each of the three approaches depends largely on context. I will constrain my discussion to the first two.

If the interpretation of an interest parameter is stable over models, it appears that first-order reliable inference via maximum likelihood estimation is possible in spite of considerable misspecification in the nuisance part of the model only when the interest parameter is orthogonal (in the sense of Jeffreys, 1948, pp. 158–184) to the notional nuisance parameters, whose

interpretation then has to be in terms of Kullback–Leibler projection. This would be a necessary condition rather than a sufficient one. Note that the true model is also implicit in the definition of parameter orthogonality. It is, as far as I am aware, an open problem to characterise the class of models whose interest and notional nuisance parameters are orthogonal under arbitrary model misspecification, perhaps after interest-respecting reparameterisation. The second-order properties are always affected, sometimes severely, which is problematic beyond point estimation. On a historical point, the limit in probability of the maximum likelihood estimator under model misspecification and its connection to the Kullback–Leibler divergence was derived by Cox (1961, 1962), who also noted the failure of Bartlett's second identity and gave a generalisation of the result (Cox, 1961, equations (28)–(43)) which later became known as the sandwich formula. A more rigorous discussion of regularity conditions was given by Huber (1967). Similar results were obtained independently by White (1982a,b).

It could be argued, contrary to the paper under discussion, that when the effects of interest are represented by parameters whose interpretations differ according to the model used, the appropriate approach is to acknowledge the model uncertainty rather than seek inference on a quantity whose interpretation is stable but perhaps only tangentially relevant when the assumed model is false. The role of sufficiency in assessment of model adequacy, implicit in R. A. Fisher's work is perhaps best approached via Barndorff-Nielsen and Cox (1994, p.29). When the ideas can be operationalised, there are no difficulties associated with double use of the data for model assessment and parametric inference. The conclusion may be that some, all or none of the a priori plausible representations are compatible with the data. If multiple models with different interpretations are not significantly contradicted, it is sometimes appropriate to report as many as feasible, a point emphasised repeatedly by D. R. Cox (e.g. Cox, 1968, 1995; Cox & Snell, 1974, p. 55; 1989, p. 193). See also Davison (1995). This underpins the development of confidence sets of models (Cox & Battey, 2017).

## REFERENCES

- Barndorff-Nielsen, O.E. & Cox, D.R. (1994) *Inference and asymptotics*. London: Chapman and Hall.
- Bartlett, M.S. (1937) Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London, Series A: Mathematical and Physical Sciences*, 160, 268–282.
- Cox, D.R. (1961) Tests of separate families of hypotheses. In: LeCam, L.M., Neyman, J. & Scott, E.L. (Eds.) *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*. Berkeley: University of California Press, 105–123.
- Cox, D.R. (1962) Further results on tests of separate families of hypotheses. *Journal of the Royal Statistical Society: Series B (Methodological)*, 24, 406–424.
- Cox, D.R. (1968) Notes on some aspects of regression analysis (with discussion). *Journal of the Royal Statistical Society: Series A*, 131, 265–279.
- Cox, D.R. (1995) Discussion of the paper by Chatfield. *Journal of the Royal Statistical Society: Series A*, 158, 455–456.
- Cox, D.R. & Battey, H.S. (2017) Large numbers of explanatory variables, a semi-descriptive analysis. *Proceedings of the National Academy of Sciences*, 114, 8592–8595.
- Cox, D.R. & Snell, E.J. (1974) The choice of variables in observational studies. *Journal of the Royal Statistical Society: Series C*, 23, 51–59.
- Cox, D.R. & Snell, E.J. (1989) *Analysis of binary data* (2nd edition). London: Chapman and Hall.
- Davison, A.C. (1995) Discussion of the paper by Chatfield. *Journal of the Royal Statistical Society: Series A*, 158, 451–452.
- Huber, P.J. (1967) The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, edited by L.M. LeCam and J. Neyman. University of California Press, Berkeley, 221–233.

- Jeffreys, H. (1948) *Theory of probability* (2nd edition). Oxford: Oxford University Press.
- White, H. (1982a) Maximum likelihood estimation of misspecified models. *Econometrica*, 50, 1–25.
- White, H. (1982b) Regularity conditions for Cox's test of non-nested hypotheses. *Journal of Econometrics*, 19, 301–318.

**How to cite this article:** Battey, H. (2022) Heather Battey's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 696–698. Available from: <https://doi.org/10.1111/rssb.12517>

DOI: 10.1111/rssb.12518

# Christian Hennig's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Christian Hennig**

University of Bologna, Bologna, Italy

## Correspondence

Christian Hennig, University of Bologna, Bologna, Italy.

Email: [christian.hennig@unibo.it](mailto:christian.hennig@unibo.it)

There is a tendency in statistics to talk about model assumptions in a misleading way. Most of us probably agree with George Box's 'all models are wrong but some are useful', yet there is much communication that implies that for applying methods 'assuming' certain models, these models have to be true. If this were so, no model-based method could ever be used!

Generally model assumptions do not have to be fulfilled. A model assumption just means that certain theoretical results regarding a statistical procedure hold assuming the model. A procedure may well deliver useful results if its model assumptions do not hold. This can be addressed by investigating what happens if other models hold. The advantage of 'assumption-lean' methods is that they come with theory that applies under a wider range of models, so we know more, but none of this wider range of models will ultimately be 'correct' either, and the theory does not necessarily guarantee a good behaviour in practice.

The authors shift the focus on estimands rather than models, although their theory and even their very definition still assumes a model, be it 'big' and non-parametric. Tukey (1997) rather

put a focus on procedures, to be challenged by studying their behaviour under different models, without assuming any of them to be true. An important issue not treated by Tukey is addressed by discussing estimands, namely whether what we estimate is what we really want to know. This requires arguments other than evidence from the data.

It may not be enough to choose an estimand so that its estimator fulfils nice asymptotic theory. The expected value  $E$  is central for the authors' estimands. Concentrating on one-dimensional location, standard robustness theory (Huber & Ronchetti, 2009) teaches that  $E$  is a non-robust functional. Looking at a model such as  $P_{\epsilon, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2} = (1 - \epsilon)\mathcal{N}(\mu_1, \sigma_1^2) + \epsilon\mathcal{N}(\mu_2, \sigma_2^2)$  with very small  $\epsilon$ , theoretically all is fine with estimating  $E$ , but  $\mathcal{N}(\mu_2, \sigma_2^2)$  may model irregular outliers, and the interest may be in estimating  $\mu_1$ . But  $EP_{\epsilon, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2}$  can be arbitrarily far away from  $\mu_1$ .

The influence function of  $E$  is unbounded (as is (13) in the paper, due to the key role of  $E$  in the estimand), which is a problem from the robustness point of view. Paraphrasing Buja et al. (2019), having a general nonparametric theory for an estimand does not necessarily imply that the estimand behaves well over the range of models covered by the theory. Whether an estimand is well chosen requires more than being 'assumption-lean'.

## REFERENCES

- Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M. et al. (2019) Models as approximations I: consequences illustrated with linear regression. *Statistical Science*, 34, 523–544 (already cited by the discussed paper).
- Huber, P.J. & Ronchetti, E.M. (2009) *Robust statistics*, 2nd edition. New York: Wiley.
- Tukey, J.W. (1997) More honest foundations for data analysis. *Journal of Statistical Planning and Inference*, 57, 21–28.

**How to cite this article:** Hennig, C. (2022) Christian Hennig's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 698–699. Available from: <https://doi.org/10.1111/rssb.12518>

The following contributions were received in writing after the meeting.



DOI: 10.1111/rssb.12519

# Pallavi Basu's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Pallavi Basu**

Indian School of Business, Hyderabad, India

## Correspondence

Pallavi Basu, Indian School of Business, Hyderabad, India.

Email: [pallavi\\_basu@isb.edu](mailto:pallavi_basu@isb.edu)

The authors propose an ingenious way to improve on providing inference for the main and interaction effects for generalised linear model parameters. With the use of the efficient influence function, they can extend the allowable bias introduced by machine learning estimators up to  $o_p(n^{-1/4})$ . Noting that cross-fitting has been suggested to counter the Donsker condition, it may be a useful alternative to contrast with machine learning estimates, averages over sample splits or even averages over several models.

The idea that the work seamlessly applies to continuous exposures without the need for conditional density estimates is indeed appealing. I am curious how this can be extended to important econometric or policy questions such as in difference-in-differences, where for dichotomous exposures, the estimand reduces to

$$E\{w(L)[Ew(L)]^{-1}[(Y^{11} - Y^{10}) - (Y^{01} - Y^{00})]\}.$$

A minimiser of the Kullback–Leibler divergence style model selection procedure of the ‘best’ model within the class of considered models has the advantage of having a model available at disposal. This may be useful in explaining to applied collaborators or providing out-of-sample estimates or prediction intervals. The philosophy of addressing model misspecification used in this work loses that connection. It could be not very clear, a priori, how to choose between the main effect versus an interaction effect estimand. Suppose there is an interest in the main effect. However, it is expected for the linear model to have interaction terms; how would one interpret the main effect estimand in the interaction model?

Although a little far-fetched, practitioners are often interested in testing the interaction effect of a group of covariates, say, the interactions of a policy and socioeconomic factors on health outcome; a model then aids in simpler joint statistical tests. Analogously, simultaneous intervals or confidence sets will be useful here. Furthermore, in the spirit of the recommended ‘hierarchical principle’, how would we interpret estimated effects when the confidence interval (CI) for the

interaction estimand does not include the null and the CI for the main effect includes zero or no effect?

Further, to connect the dots well, it will be great to understand the causal identifying conditions, considering that inference rather than prediction is the goal, an edge that will greatly help declutter the uses of black-box methodologies. Thereafter, connecting to unmeasured confounding either via Rosenbaum's sensitivity or VanderWeele's  $E$ -value will be beneficial. However, the inferential nature of the causal estimand, either testing or providing valid confidence intervals, should remain.

**How to cite this article:** Basu, P. (2022) Pallavi Basu's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 700–701. Available from: <https://doi.org/10.1111/rssb.12519>

DOI: 10.1111/rssb.12520

# Blair Bilodeau's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

## Blair Bilodeau

University of Toronto, Toronto, Ontario, Canada

### Correspondence

Blair Bilodeau, University of Toronto, Toronto, Ontario, Canada.

Email: [blair.bilodeau@mail.utoronto.ca](mailto:blair.bilodeau@mail.utoronto.ca)

I commend Vansteelandt and Dukes for (a) advocating that statistics move beyond well-specified assumptions and (b) providing a practical estimator to address this for generalised linear models (GLMs). Furthermore, their derivation of theoretical guarantees for this estimator is highly attractive. However, the theorem assumptions exclude many standard machine learning methods, which may violate the spirit of 'assumption-lean' inference guarantees.

Theorems 2 and 4 assume that the three (non-parametric) conditional mean estimates have expected square loss converging at rate *strictly* faster than  $n^{-1/2}$ . Theorem 7 of Rakhlin et al. (2017) can be extended to show that if the empirical  $L_2$  entropy of the model class depends on the scale  $\varepsilon \in (0, 1)$  at rate  $\varepsilon^{-p}$  for some  $p > 0$ , then the minimax expected square loss is lower bounded

by  $n^{-2/(p+2)}$  even in the well-specified setting. That is, in order for the theory of Vansteelandt and Dukes to apply without additional assumptions, the three conditional mean models *must* each be strictly Donsker ( $p < 2$ ), and this cannot be readily side-stepped using sample splitting as Vansteelandt and Dukes do elsewhere.

The strict Donsker assumption is satisfied by models with the number of ‘parameters’ (e.g. weights of a neural network) growing strictly slower than  $n$ . However, a fixed number of parameters is exactly a parametric assumption, and in practice the number of parameters in machine learning models is often larger than  $n$  (corresponding to the *interpolation* regime, see Belkin et al., 2019). In the interpolation regime, such models are better understood from a non-parametric perspective; unfortunately, this means that they suffer from an exponential dependence on the dimension of the inputs.

For linear models, the dimension-free entropy growth rate is  $\varepsilon^{-2}$  (Mendelson & Schechtman, 2004; Zhang, 2002), and consequently, the present theory does not apply. For integer  $\alpha$  and  $\gamma \in (0, 1]$ , entropy for the class of  $(\alpha, \gamma)$ -Hölder smooth functions on  $\mathbb{R}^d$  grows at rate  $\varepsilon^{-d/(\alpha+\gamma)}$  (Wainwright, 2019, Example 5.11), requiring univariate inputs to apply the present theory in the Lipschitz ( $\alpha = 0, \gamma = 1$ ) setting. These examples include neural networks with linear and Lipschitz activations, as well as certain variants of random forests (Mourtada et al., 2020). In the notation of Vansteelandt and Dukes, the input dimension corresponds to the dimension of  $A$  and  $L$  jointly, and in practice  $L$  can be quite high dimensional.

The authors clearly acknowledge that their convergence requirements may not be satisfied by ‘very flexible machine learning methods’. However, these requirements exclude many methods of interest, including standard neural network and random forest architectures. Ultimately, it remains an interesting open question whether Vansteelandt and Dukes’ estimator enjoys convergence guarantees (even with appropriately slow rates) when used with such canonical nonparametric methods.

## REFERENCES

- Belkin, M., Hsu, D., Ma, S. & Mandal, S. (2019) Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences of the United States of America*, 116(32), 15849–15854. <https://doi.org/10.1073/pnas.1903070116>
- Mendelson, S. & Schechtman, G. (2004) The shattering dimension of sets of linear functionals. *Annals of Probability*, 32(3A), 1746–1770.
- Mourtada, J., Gaïffas, S. & Scornet, E. (2020) Minimax optimal rates for Mondrian trees and forests. *The Annals of Statistics*, 48(4), 2253–2276. <https://doi.org/10.1214/19-AOS1886>
- Rakhlin, A., Sridharan, K. & Tsybakov, A.B. (2017) Empirical entropy, minimax regret and minimax risk. *Bernoulli*, 23(2), 789–824. <https://doi.org/10.3150/14-BEJ679>
- Wainwright, M.J. (2019) *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press.
- Zhang, T. (2002) Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2, 527–550.

**How to cite this article:** Bilodeau, B. (2022) Blair Bilodeau’s contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’ by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 701–702. Available from: <https://doi.org/10.1111/rssb.12520>

DOI: 10.1111/rssb.12521

# Andreas Buja, Richard A. Berk, Arun K. Kuchibhotla, Linda Zhao and Ed George’s contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’ by Vansteelandt and Dukes

Andreas Buja<sup>1,2</sup> | Richard A. Berk<sup>3</sup> | Arun K. Kuchibhotla<sup>4</sup> | Linda Zhao<sup>1</sup> | Ed George<sup>1</sup>

<sup>1</sup>Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>2</sup>Flatiron Institute, Simons Foundation, New York, USA

<sup>3</sup>Department of Criminology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>4</sup>Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

## Correspondence

Andreas Buja, Flatiron Institute, Simons Foundation, 162 Fifth Avenue, New York, NY 10010, USA.

Email: [abuja@flatironinstitute.org](mailto:abuja@flatironinstitute.org)

We thank the authors for their stimulating article. At the highest level, it agrees with our thinking as expressed in the second of two articles in *Statistical Science* (Buja et al., 2019). The guiding principle is to focus on quantities of interest that are meaningful statistical functionals, beyond the statistical models that motivate them, even when those models are fully misspecified. In some form, this focus has existed for decades in the semi-parametric literature. We suggest below a reinterpretation of the authors’ approach and a modification of their third principle.

**1. Interpretation:** The authors’ formulation can be reinterpreted as follows:

- Dispense with the complexities of GLMs, which for a mixed semi-parametric model, ignoring scale and normalization, are:

$$p(y:\theta) \sim \exp(y\theta - b(\theta)), \theta = \beta A + \omega(L), g^{-1}(\theta) = b'(\theta), E[Y | A, L] = g^{-1}(\theta)$$

Ignore also ‘proper’ minimization of the negative log-likelihood:

$$\min_{\beta, \omega(\cdot)} E[b(\beta A + \omega(L)) - Y \cdot (\beta A + \omega(L))].$$

- Instead, use OLS to achieve  $g(\mathbf{E}[Y | A, L]) \approx \theta = \beta A + \omega(L)$  directly:

$$\min_{\beta, \omega(L)} \mathbf{E} \left[ \left( g(\mathbf{E}[Y | A, L]) - (\beta A + \omega(L)) \right)^2 \right] \quad (1)$$

For a binary outcome  $Y$  and logistic regression, this means applying OLS to the logit of  $\mathbf{E}[Y | A, L]$ .

The authors' proposal (5) is the solution to the minimization problem (1), which shows that it is a projection functional based on a non-standard loss function.

The parameter of interest,  $\beta$ , is obtained by adjusting  $g(\mathbf{E}[Y | A, L])$  and  $A$  non-parametrically for  $L$ ,

$$Z := g(\mathbf{E}[Y | A, L]) - \mathbf{E}[g(\mathbf{E}[Y | A, L]) | L] \text{ and } X := A - \mathbf{E}[A | L]$$

and applying a simple linear regression to  $Z$  vs.  $X$ :  $\beta = \mathbf{E}[ZX] / \mathbf{E}[X^2]$ .

**2. Suggestion:** The authors' third principle states that the parametric estimand should be an  $L$ -weighted average of estimands at each stratum of the confounder  $L$ . However, the estimands are not covariances but slopes,

$$\beta | L = \frac{\text{Cov}(g(\mathbf{E}[Y | A, L]), A | L)}{\text{Var}(A | L)},$$

for which averaging over  $L$ -strata is incorrect. There is no need for a principle if the quantity of interest is mathematically defined as a functional on distributions, be they cofounder-conditional ( $\beta | L = \mathbf{E}[ZX | L] / \mathbf{E}[X^2 | L]$ ) or marginal ( $\beta = \mathbf{E}[ZX] / \mathbf{E}[X^2]$ ). Yet, there is value in the idea of weighting across  $L$ -strata because it can be used to generate multiple functionals that satisfy the authors' first principle: agreement with the model parameter if the model is correct. This works as follows:

- Reweight the data/population with an  $L$ -dependent weight function  $w(L)$  (not to be confused with  $\omega(L)$  of the model):

$$p_w(y, a, l) = w(l) p(y, a, l), \quad \text{where } \mathbf{E}_p[w(L)] = 1.$$

- Apply the functional of interest to the reweighted data/population:

$$\beta_w = \frac{\mathbf{E}[w(L) \cdot \text{Cov}(g(\mathbf{E}[Y | A, L]), A | L)]}{\mathbf{E}[w(L) \cdot \text{Var}(A | L)]} = \frac{\mathbf{E}[w(L) ZX]}{\mathbf{E}[w(L) X^2]}.$$

This specializes to the authors' proposal (5) for  $w(L) \equiv 1$ .

When the model is 1<sup>st</sup> order correct,  $g(\mathbf{E}[Y | A, L]) = \beta A + \omega(L)$ , it holds that  $\beta_w$  is the same for all weight functions  $w(L)$ :  $\beta_w \equiv \beta$ . When the model is 1<sup>st</sup> order misspecified, there exist weight functions  $w_1(L)$  and  $w_2(L)$  such that  $\beta_{w_1} \neq \beta_{w_2}$ , which is equivalent to 'effect heterogeneity', i.e.,

$\beta \mid L$  is **not** the same (a.s.) for all confounder strata  $L$ . This lends itself to a ‘well-specification’ diagnostic for regression functionals, as elaborated in Buja et al. (2019). It specializes here to a diagnostic for effect heterogeneity.

## REFERENCE

Buja, A., Berk, R., Brown, L., George, E., Kuchibhotla, A.K. & Zhao, L. (2019) Models as approximations, part II: a general theory of model-robust regression. *Statistical Science*, 34(4), 545–565.

**How to cite this article:** Buja, A., Berk, R.A., Kuchibhotla, A.K., Zhao, L. & George, E. (2022) Andreas Buja, Richard A. Berk, Arun K. Kuchibhotla, Linda Zhao and Ed George’s contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’ by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 703–705. Available from: <https://doi.org/10.1111/rssb.12521>

DOI: 10.1111/rssb.12522

# Anna Choi and Weng Kee Wong’s contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’ by Vansteelandt and Dukes

Anna Choi<sup>1</sup> | Weng Kee Wong<sup>2</sup>

<sup>1</sup>Stanford University, Stanford, California, USA

<sup>2</sup>UCLA, Los Angeles, California, USA

## Correspondence

Weng Kee Wong, UCLA, Los Angeles, California, USA

Email: [wk Wong@ucla.edu](mailto:wk Wong@ucla.edu)

We congratulate the authors for major advances in the methodology of estimands, in particular assumption-lean reference for the estimands in possibly misspecified generalised linear models. Whereas the simulation studies in Section 6 indeed support the ‘assumption-lean’ claim and the authors’ software in *Practice of Epidemiology* **187** (pp. 1079–1084, 2018) facilitates such inference, the empirical example in Section 7 seems problematic to illustrate applications of the methodology. The example focuses on ‘data obtained from birth certificates from 2500 children born in King County, Washington in 2001’ in the First Steps program that was ‘set up in 1989’

in King County. These are longitudinal data, and the program is still ongoing. The methodology for longitudinal studies developed by Laird and Ware (1982) and applied to the Six Cities Study (Ryan, 2015, section 3) using generalised linear mixed models seems more appropriate for the re-analysis of the longitudinal data from the First Steps program of King County. Although Laird and Ware's methodology does not involve causal inference, Laird mentions how such inference can be added to the analysis in Ryan (2015, section 9 on statistical genetics). Chen et al. (2018, sections 3.6.3, 4.5, 5.4, 6.3, 6.4, 6.5, 7.4, 7.5) describe how causal inference/conclusions can be derived via multi-criteria statistical decision theory, exposure-adjusted incidence rates and multiple testing with familywise error or false discovery rate control for clinical trials data, control for confounding using inverse probability weighting, propensity scores, instrumental variables and research designs for unmeasured confounders, structural causal models and symbolic causal calculus for post-marketing safety data, empirical Bayes and Bayesian approaches to signal detection from adverse event databases. In addition, Sections 2.1.3 and 2.1.4 of that book describe recent advances in statistical learning methods which involve nonlinear basis functions such as neural networks and classification/regression trees and in validation of the prediction/regression model, in contrast to the linear (or partial linear) basis functions used in the present paper and complementary to the sample splitting asymptotic theory in its Theorem 2.

## REFERENCES

- Chen, J., Heyes, J. & Lai, T.L. (2018) *Medical product safety evaluation: biological models and statistical methods*. Boca Raton FL: Chapel R Hall/CRC.
- Laird, N. & Ware, J. (1982) Random-effects models for longitudinal data. *Biometrics*, 38, 963–974.
- Ryan, L. (2015) A conversation with Nan Laird. *Statistical Science*, 30(4), 582–596.

**How to cite this article:** Choi, A. & Kee Wong, W. (2022) Anna Choi and Weng Kee Wong's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 705–706. Available from: <https://doi.org/10.1111/rssb.12522>

DOI: 10.1111/rssb.12523

# Chaohua Dong, Jiti Gao and Oliver Linton's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Chaohua Dong<sup>1</sup> | Jiti Gao<sup>2</sup> | Oliver Linton<sup>3</sup>

<sup>1</sup>Zhongnan University of Economics and Law, Wuhan, China

<sup>2</sup>Monash University, Melbourne, Australia

<sup>3</sup>University of Cambridge, Cambridge, UK

## Correspondence

Oliver Linton, University of Cambridge, Cambridge, UK.

Email: [obl20@outlook.com](mailto:obl20@outlook.com)

The title of this paper is ironically self-fulfilling, since there are almost no meaningful assumptions made throughout! The starting point is that we have some plausible semi-parametric model, which is a special case of a more general non-parametric model, but we wish to allow for misspecification and in particular define an estimand that is meaningful in the non-parametric model and that specialises in the plausible model to a slope coefficient. However, since the estimator that is proposed is not the semi-parametric efficient estimator of that slope coefficient under the semi-parametric model, we wonder what is the role of the model at all? The theory side of it seems to assume in Theorem 2, for example that  $E(Y|A, L)$  is consistently estimated in  $L_2$  under the full unrestricted -parametric setting. But if that is possible, then why bother with the model? The authors talk casually about machine learning methods being used to estimate  $E(Y|A, L)$ , but if that is a silver bullet, then who needs the model? The model embodies some structure around  $A, L$  but the discussion is focussed away from the dimensionality of  $L$ , which is a big reason why one might want a structured model such as additivity (Linton & Nielsen, 1995) or the partial linear model (Robinson, 1988). Perhaps it would help if a full model was written down for the effect of high-dimensional  $L$ . Perhaps the point is that the parameter of interest is only defined in terms of low-dimensional conditional expectations, but this does not appear to be the case in the sense that high-dimensional smoothing is employed in (a) of p.14, which is then projected down by conditional expectation onto  $L$ , but if  $A$  is binary, then this has not reduced dimensionality at all, the dimensionality issue sits in  $L$  and what structure is assumed about its effect on  $Y$ .

---

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.



The authors say that the usual choice of smoothing parameter is not tuned to the estimand. This point has been made in Linton (1995) who derived an optimal bandwidth for estimation of slope parameters and Wald statistics in the partially linear model based on local polynomial estimators; the optimal rates are indeed different from those that minimise the mean squared error of the non-parametric regression involved.

An alternative approach is to use sieve methods throughout and penalisation. Of course this questions whether it is necessary to pay too much attention to the approximating model. For example, suppose that we let  $X_i$  be the  $(2dK + 1) \times 1$  vector containing  $A_i$  and basis functions  $\psi_k(L_{ji})$  and  $A_i\psi_k(L_{ji})$  (if interactions between  $L_j$  and  $A$  are important) for  $k = 1, \dots, K$  and  $j = 1, \dots, d$ , and  $i = 1, \dots, n$ , and let  $\theta \in \mathbb{R}^{2dK+1}$  minimise

$$\left\| \frac{1}{n} \sum_{i=1}^n m(Y_i, \theta^\top X_i) \right\| + \text{pen}_\lambda(\theta), \quad (1)$$

where  $m$  is a large vector of (possibly non-linear) moment condition, while  $\text{pen}_\lambda$  is a penalty function such as SCAD or LASSO. Dong et al. (2018) establish, as a special case, the consistency and asymptotic normality of the estimators in (1) and provide consistent inference methods when the dimensionality of  $X_i$  is diverging and a smooth penalty like SCAD is used.

## REFERENCES

- Dong, C., Gao, J. & Linton, O. (2018) High dimensional semiparametric moment restriction models. To Appear in *Journal of Econometrics*. Available from: <https://core.ac.uk/reader/186326795>
- Linton, O. (1995) Second order approximation in the partially linear regression model. *Econometrica*, 63, 1079–1112.
- Linton, O. & Nielsen, J. (1995) A kernel method of estimating nonparametric structured regression based on marginal integration. *Biometrika*, 82, 93–100.
- Robinson, P.M. (1988) Root-N-consistent semiparametric regression. *Econometrica*, 56, 931–954.

**How to cite this article:** Dong, C., Gao, J. & Linton, O. (2022) Chaohua Dong, Jiti Gao and Oliver Linton's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 707–708. Available from: <https://doi.org/10.1111/rssb.12523>

DOI: 10.1111/rssb.12524

# Oliver Hines and Karla Diaz-Ordaz's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Oliver Hines | Karla Diaz-Ordaz

Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK

## Correspondence

Oliver Hines, Department of Medical Statistics, London School of Hygiene and Tropical Medicine, London, UK.

Email: [oliver.hines@lshtm.ac.uk](mailto:oliver.hines@lshtm.ac.uk)

We congratulate the authors on their work, which contributes to a growing movement in Statistics, moving away from (semi)parametric-based regression model assumptions, in favour of understanding working models as non-parametric projections of the true distribution. Here, we highlight this aspect of the work, and draw connections with other recent developments.

Consider the functional  $m(A, L) = g\{E(Y|A, L)\}$ , using the same notation as Vansteelandt and Dukes (2020). Without loss of generality,

$$m(A, L) = \beta A + \omega(L) + R(A, L),$$

where  $\beta$  is an arbitrary constant,  $\omega(\cdot)$  is an arbitrary function and  $R(A, L)$  is a remainder term, without restrictions.

The *parametric modeller* assumes that for some particular  $\beta$ ,  $\omega(\cdot)$ , the remainder term is exactly zero. Inference is usually carried out under this assumption, resulting in the dishonest inference described in (Vansteelandt & Dukes, 2020). The *projectionist*, however, instead aims to report  $\beta$  such that the remainder term is, in some sense, small. The projectionist argues that when the remainder is small,  $\beta$  captures the main effect of  $A$  on  $Y$ . Indeed when the parametric modeller is correct, both the reported main effects coincide.

One sense in which the remainder might be small is in squared expectation, e.g.  $\beta$  is

$$\arg \min_{\beta \in \mathbb{R}} E[\{R(A, L)\}^2] = \arg \min_{\beta \in \mathbb{R}} E[\{m(A, L) - \beta A - \omega(L)\}^2] = \frac{E[\text{cov}\{A, m(A, L)|L\}]}{E\{\text{var}(A|L)\}},$$

where  $\omega(\cdot)$  is chosen to minimise the same quantity. This minimisation recovers the main effect of Vansteelandt and Dukes (2020), which is consequently given a least squares projection interpretation.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

The estimand, however, remains non-parametrically defined and can be estimated using methods described in (Vansteelandt & Dukes, 2020).

Vansteelandt and Dukes (2020) also elucidates the relationship between causally-derived estimands and those defined through parametric models, which are seemingly disparate. Consider, for example, a conditional effect estimand,  $\lambda(\cdot)$ , defined by

$$m(A, L) = \lambda(L)A + \omega(L) + R(A, L).$$

$\lambda(\cdot)$  can be interpreted as a heterogeneous treatment effect (HTE). To see why, notice that for an identity link, and binary  $A$ , the remainder is zero when  $\lambda(L)$  is the HTE and  $\omega(L)$  is the conditional mean outcome in the  $A = 0$  treatment subgroup, i.e.  $m(A, L)$  is written

$$E(Y|A, L) = \{E(Y|A = 1, L) - E(Y|A = 0, L)\}A + E(Y|A = 0, L).$$

Regardless of whether  $A$  is binary or not, the projectionist may decide to report  $\lambda(\cdot)$ , such that the mean squared remainder is minimised

$$\lambda(\cdot) = \arg \min_{\lambda(\cdot) \in \mathcal{F}} E[\{m(A, L) - \lambda(L)A - \omega(L)\}^2] = \frac{\text{cov}\{A, m(A, L)|L=\cdot\}}{\text{var}(A|L=\cdot)},$$

where  $\mathcal{F}$  is the set of functions mapping confounder vectors to the real numbers, and  $\omega(\cdot)$  minimises the same objective function. When  $A$  is binary and  $g(\cdot)$  the identity link,  $\lambda(\cdot)$  reduces to the binary HTE; however, it remains well defined for continuous and discrete exposures. Moreover, we note that the recent R-learner (Nie & Wager, 2021; Robinson, 1988) provides a method for estimating  $\lambda(\cdot)$  when  $g(\cdot)$  is the identity link, since the R-learner minimises the objective function,

$$\arg \min_{\lambda(\cdot) \in \mathcal{F}} E[(Y - E(Y|L) - \lambda(L)\{A - E(A|L)\})^2] = \frac{\text{cov}(A, Y|L=\cdot)}{\text{var}(A|L=\cdot)}.$$

Finally we observe that the main effect estimand in (Vansteelandt & Dukes, 2020) is a weighted average of this HTE,

$$E \left\{ W(L) \frac{\text{cov}\{A, m(A, L)|L\}}{\text{var}(A|L)} \right\},$$

with the weight  $W(L) = \text{var}(A|L)/E\{\text{var}(A|L)\}$ , which is non-negative and has expectation 1.

We believe that model projection estimands represent an exciting frontier in statistical research.

## REFERENCES

- Nie, X. & Wager, S. (2021) Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2), 299–319.  
 Robinson, P.M. (1988) Root-N-consistent semiparametric regression. *Econometrica*, 56(4), 931.  
 Vansteelandt, S. & Dukes, O. (2020) Assumption-lean inference for generalised linear model parameters. *arXiv*.  
 Available from: <https://doi.org/10.48550/arxiv.2006.08402>

**How to cite this article:** Hines, O. & Diaz-Ordaz, K. (2022) Oliver Hines and Karla Diaz-Ordaz's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 709–710. Available from: <https://doi.org/10.1111/rssb.12524>

DOI: 10.1111/rssb.12525

# Ian Hunt's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Ian Hunt**

University of Tasmania, Hobart, TAS, Australia

**Correspondence**

Ian Hunt, University of Tasmania, Hobart TAS, Australia.

Email: [ihunt@bunhill.co.uk](mailto:ihunt@bunhill.co.uk)

This paper offers a generic and principled addition to conventional statistical modelling. It is a real advance in *applicable* methodology. But there is no need in the paper for the following claims: that statisticians routinely use 'dishonest' modelling assumptions, that statisticians should aim for 'purely evidence-based' inferences and that modelling assumptions are 'almost always a pure mathematical convenience'. These sorts of claims are redundant or misleading for three reasons.

First, text-book statistical inferences look like deductions. But nearly everyone wants to make inductive inferences. Statisticians assist in the leap from deduction to induction via *iterations* of a two-step process: step one, specify different models and estimate their parameters (adding and dropping interaction terms, using different error assumptions and so on); step two, assess and compare the models (including assumption checking). This inductive process is necessarily vague, but it embodies the key statistical tasks defined by Fisher (1959, pp. 6–8). Properly disclosed assumptions that cannot be proved, or are openly false, play an *honest* and essential role in inductive processes—we have to take risks to go beyond the data and what we already know. Adding assumption-lean or 'less risky' methods to inductive processes may prove useful but this would not be 'more honest'.

Second, if by 'evidence' the authors really mean data then I would argue that aiming for 'purely evidence-based' inferences is not ambitious enough. Seeking inductive or causal inferences is to aim higher, but 'no causes in, no causes out' (Cartwright, 1994) and, I argue, 'no assumptions in, no inductions out'. In other words, substantive and risky modelling assumptions help turn data into the sort of evidence that supports causal and inductive claims.

Third, I acknowledge the convenience-factor of conventional modelling assumptions; proverbial examples include simplistic error structures and the absence of model mis-specification. But with finite samples, the methodology promoted in this paper has its own hazards tucked away. For example, estimators may depend on the particular machine learning method used (kitchen-sink approaches, like SuperLearner, are no panacea); and sample-splitting or cross-fitting can cause biases (such as a tendency to over-penalize more complex models that require larger samples for adequate parameter estimation) as well as noise.

---

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Author. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

Assumption-lean methods are not necessarily superior to assumption-rich methods, but they are surely complimentary. In future applied work, I look forward to *adding* some of these generic estimand-centric methods to my usual, honestly-reported, model comparisons and tentative inductive leaps.

## REFERENCES

- Cartwright, N. (1994) *Nature's capacities and their measurement*. Ch2. Oxford: Oxford University Press.
- Fisher, R.A. (1959) *Statistical methods, experimental design and scientific inference*, 1990 edition. Oxford: Oxford University Press.

**How to cite this article:** Hunt, I. (2022) Ian Hunt's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 711–712. Available from: <https://doi.org/10.1111/rssb.12525>

DOI: 10.1111/rssb.12526

# Kuldeep Kumar's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

## Kuldeep Kumar

Bond University, Gold Coast, Queensland, Australia

### Correspondence

Kuldeep Kumar, Bond University, Gold Coast, Queensland, Australia.

Email: [kkumar@bond.edu.au](mailto:kkumar@bond.edu.au)

According to Leo Breiman (2001), there are two broad cultures for analysing and modelling to reach conclusions from the data. The first one is data modelling culture where the value of the parameters are estimated from the data and then the model is used for information and/or prediction. The second one is algorithmic modelling culture, where the approach is to find a function  $f(x)$  using an algorithm that operates on  $x$  to predict  $y$ . In the first case, statistical tools like OLS are used whereas machine learning tools perform much better in the second case. I should

---

This is an open access article under the terms of the [Creative Commons Attribution](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Author. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

congratulate the authors for developing a hybrid model using machine learning and traditional statistical techniques like OLS. However, I do have few queries on the real-life applications of the proposed model. Most of the simulation results for main effects, effect modification and misspecified link function are valid for sample size  $n = 500$  or more. Also, the data analysis example presented in section 7 has a sample size of 2500. How is the empirical bias affected when the sample size is small? Bayesian models incorporate prior information, which can be specified quantitatively in the form of a distribution. Do the authors think a Bayesian parameter generalised linear model can perform better if some prior information is available? Also, I am not sure how this model tackles the problem of multicollinearity? Finally, in the context of high dimensional variable selection, what will happen if the covariates are not distributed normally?

## REFERENCE

Breiman, L. (2001) Statistical modeling: the two cultures. *Statistical Science*, 16(3), 199–231.

**How to cite this article:** Kumar, K. (2022) Kuldeep Kumar's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 712–713. Available from: <https://doi.org/10.1111/rssb.12526>

DOI: 10.1111/rssb.12527

# Michael Lavine and James Hodges' contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Michael Lavine<sup>1</sup> | James Hodges<sup>2</sup>

<sup>1</sup>Department of Mathematics and Statistics, University of Massachusetts, Amherst, Massachusetts, USA

<sup>2</sup>Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota, USA

## Correspondence

James Hodges, Division of Biostatistics, University of Minnesota, Minneapolis, MN, USA.

Email: [hodge003@umn.edu](mailto:hodge003@umn.edu)

The authors advocate a style and rhetoric of statistical analysis that begins with specifying an estimand and ends with an estimate and interval. They say, 'The starting point... is to come up

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

with an estimand that is meaningful when the above [generalized partially linear] model [(4)] does not hold, but reduces to [the familiar regression coefficient]  $\beta$  when the model holds; this... allows for nonparametric inference to be developed for that estimand'. However, in our opinion, such an analysis is not always appropriate. For example, consider the second of Anscombe's (1973) data sets, in which  $Y$  is a quadratic function of  $X$ . The authors'  $Y$  is Anscombe's  $Y$ ; the authors'  $A$  is Anscombe's  $X$ ; and the authors'  $L$  has just one value. The authors' estimand appears to be given by (5), where  $g$  is the identity function. For a given data set, including Anscombe's, (5) reduces to one number which is to be interpreted similarly to a slope. In our opinion, that is a poor way to summarize Anscombe's data. The problem arises because the authors insist on choosing an estimand before seeing the data, thereby precluding the possibility of summarizing the data with a statement like ' $E[Y|X]$  appears to be a roughly unimodal function of  $X$  with a maximum around  $X = 11$ '. Contrary to the authors' claim that 'the resulting analysis can be pre-specified, which is essential if one aims for an honest data analysis that reflects all uncertainties', we think it is dishonest, or at least a serious mistake, to summarize Anscombe's data with one slope-like number.

Instead, we prefer a style and rhetoric of sensitivity, which begins with looking at the data to see what sorts of models provide useful descriptions; then uses familiar, interpretable quantities and a simple analysis; and then considers how substantive conclusions change with deviations from the assumptions of the simple analysis, elaborated only as needed. The product is a leading result and a summary of the effects of variations. Barr et al. (2012) is but one of many possible examples of this style and rhetoric, the regression diagnostics literature being another. From our experience, a rhetoric of sensitivity provides a more intelligible and better-fitted answer to our scientific collaborators' questions and concerns, at least for the range of problems we have worked, as well as being more transparent.

## REFERENCES

- Anscombe, F.J. (1973) Graphs in statistical analysis. *American Statistician*, 27(1), 17–21.
- Barr, C.D., Diez, D.M., Wang, Y., Dominici, F. & Samet, J.M. (2012) Comprehensive smoking bans and acute myocardial infarction among Medicare enrollees in 387 US counties: 1999–2008. *American Journal of Epidemiology*, 176(7), 642–648.

**How to cite this article:** Lavine, M. & Hodges, J. (2022) Michael Lavine and James Hodges' contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 713–714. Available from: <https://doi.org/10.1111/rssb.12527>

DOI: 10.1111/rssb.12528

# Elizabeth L. Ogburn, Junhui Cai, Arun K. Kuchibhotla, Richard A. Berk and Andreas Buja's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Elizabeth L. Ogburn<sup>1</sup> | Junhui Cai<sup>2</sup> | Arun K. Kuchibhotla<sup>3</sup> |  
Richard A. Berk<sup>2,4</sup> | Andreas Buja<sup>5</sup>

<sup>1</sup>Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA

<sup>2</sup>Department of Statistics and Data Science, The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>3</sup>Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

<sup>4</sup>Department of Criminology, University of Pennsylvania, Philadelphia, Pennsylvania, USA

<sup>5</sup>Flatiron Institute, Simons Foundation, New York, New York, USA

## Correspondence

Elizabeth L. Ogburn, John Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

Email: [eogburn@jhsph.edu](mailto:eogburn@jhsph.edu)

Not all conditional associations between outcomes and exposures are of interest. Those that are tend to be directional: up or down. The simplest way to assess directionality is to fit a confounder-adjusted linear exposure term, as the authors propose. We agree with this approach as some of us have argued that linear slopes are meaningful and interpretable even if the directional association is not linear (Buja et al., 2019, section 10). The authors, and Whitney et al. (2019), remind us that severely misspecified adjustment can result in distortions of linear exposure slopes. In their examples, the  $A-L$  distributions have U-shaped nonlinearities and, as a result, naive linear adjustment produces a biased estimate of the true slope. Thorough data analysis could unearth such exposure-confounder structure if present in real data. A greater worry for practitioners is missing an essential confounder that biases or reverses the direction of association. The authors' inferential framework does not require  $L$  to control for all  $A-Y$  confounding, but meaningful use

---

We congratulate the authors on their excellent article (Vansteelandt and Dukes, 2021). In this comment, we highlight a few practical issues related to their proposal.

---

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](#) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.



of the estimand likely does—and therefore practitioners should select  $L$  with care and interpret estimates in conjunction with sensitivity analyses.

The authors' project of assumption lean inference rests on the assumption that nuisance parameters can be estimated nonparametrically at rate  $n^{1/4}$ . It is surprising to us that this property is widely assumed to hold for machine learning methods. The authors are in good company with this assumption, but, for example, the random forests included in the authors' analyses can have large bias if a tuning parameter is chosen badly (Olson, 2018), and as far as we know cross-validation has not been shown to reliably choose good tuning parameters. Even if  $n^{1/4}$  rates are achieved asymptotically, slower rates of convergence may require large samples before asymptotic approximations are useful. This points to the importance of methods to test or help ensure that the required rates are achieved (Liu et al., 2020; Robins et al., 2008; van der Laan et al., 2021), or to perform valid inference under slower rates (Cattaneo and Jansson, 2018; Kuchibhotla et al., 2021).

We re-ran the authors' code and applied HulC, a new method for the construction of assumption—lean confidence intervals (Kuchibhotla et al., 2021).<sup>1</sup> We found that the point estimates are indeed sensitive to choice of tuning parameters. Although HulC intervals are wider, they are valid even if approximate normality does not hold, as would be the case if the nuisance estimators converge slower than  $n^{-1/4}$ , as long as the estimator satisfies a weaker median unbiasedness property (Kuchibhotla et al., 2021).

**How to cite this article:** Ogburn, E.L., Cai, J., Kuchibhotla, A.K., Berk, R.A. & Buja, A. (2022) Elizabeth L. Ogburn, Junhui Cai, Arun K. Kuchibhotla, Richard A. Berk and Andreas Buja's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 715–716. Available from: <https://doi.org/10.1111/rssb.12528>

<sup>1</sup>The data and code were provided by the authors. We modified the code slightly, removing the support vector machine method from the SuperLearner library because of an error message. Because of this, our point estimates are close, but not identical, to those reported by the authors. The code to produce all tables is available at <https://github.com/cccfan/HulC-on-VD>.

DOI: 10.1111/rssb.12529

# Rachael V. Phillips and Mark J. van der Laan's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Rachael V. Phillips  | Mark J. van der Laan 

Division of Biostatistics, University of California, Berkeley, California, USA

## Correspondence

Mark J. van der Laan, University of California, Berkeley, CA, USA.

Email: [laan@berkeley.edu](mailto:laan@berkeley.edu)

We commend the authors on this novel non-parametric extension of main and interaction coefficients in a generalised linear model (GLM), such as  $\mathbb{E}(Y|A, L) = \beta A + g(L)$ , and their efficient estimation. There is a wealth of non-parametric extensions one could pursue, including weighted squared error projections of a conditional treatment effect  $\mathbb{E}(Y|A = a, L) - \mathbb{E}(Y|A = 0, L)$  and conditional interaction effect  $\mathbb{E}(Y|A = (a_1 + 1, a_2 + 1), L) - \mathbb{E}(Y|A = (a_1, a_2 + 1), L) - \mathbb{E}(Y|A = (a_1 + 1, a_2), L) + \mathbb{E}(Y|A = (a_1, a_2), L)$  onto a simple parametric models  $\beta a$  and  $\beta_0 + \beta_1 a_1 + \beta_2 a_2$ , respectively (Chambaz et al. (2012)). Such projection estimands are common in the causal inference literature. The authors propose a different estimand criterion that requires the efficient influence curve (EIC) to avoid (conditional) density estimation and inverse weighting. Inverse weighting can certainly cause instability but machine learning techniques have been well-adapted for conditional density estimation (for instance, Muñoz and van der Laan (2011), Rytgaard et al. (2021) and van der Laan (2010)). The above least squares projection for the main effect has an EIC that inverse weights by  $P(A = 0|L)$ , thereby achieving the first but not the second goal.

The proposed main effect estimand  $\mathbb{E}((A - \mathbb{E}(A|L))(\mathbb{E}(Y|A, L) - \mathbb{E}(Y|L)))/\mathbb{E}\sigma_{A|L}^2$  (5) satisfies this criteria for the EIC; however, it is harder to interpret outside the GLM. For example, if  $(\mathbb{E}(Y|A, L) - \mathbb{E}(Y|L))$  is not linear in  $(A - \mathbb{E}(A|L))$ , then the numerator will generally average both negative and positive contributions  $\mathbb{E}(Y|A, L) - \mathbb{E}(Y|L)$ , even for problems in which  $\mathbb{E}(Y|A, L) - \mathbb{E}(Y|A = 0, L) \geq 0$  everywhere. Consider  $A \sim U(0, 1)$  independent of  $L$  and  $\mathbb{E}(Y|A, L) = \epsilon^{-1}A\mathbb{I}(A \leq \epsilon) + \mathbb{I}(A > \epsilon) + g(L)$ . Here, the numerator becomes  $\mathbb{E}(A - 1/2)(\epsilon^{-1}A\mathbb{I}(A \leq \epsilon) + \mathbb{I}(A > \epsilon))$  and approximates  $\mathbb{E}(A - 1/2) = 0$  as  $\epsilon \rightarrow 0$ , whereas the unweighted projection of  $\mathbb{E}(Y|A, L) - \mathbb{E}(Y|A = 0, L)$  on  $\beta a$ , would result in  $\beta = 3/2 - \epsilon^2/2$ , correctly demonstrating a strong treatment effect. In addition to impacting the interpretation, these cancellations can hurt the power for testing a null hypothesis relative to testing with the parameter in Chambaz et al. (2012), even though the latter's EIC might have larger variance. The interpretation of the interaction estimand (e.g., 10) for continuous  $A_1, A_2$  has an additional complication in the sense that it does not involve an average of  $L$ -specific interactions.

Both of the proposed estimands and their EICs depend on the conditional distribution of  $A$ , given  $L$ , making the interpretations non-robust to irrelevant deviations in the study. Users should carefully consider knowledge regarding the model for  $P(A|L)$  and compute the EIC accordingly, potentially achieving efficiency gains relative to the non-parametric model considered.

Finally, the authors proposed a one-step estimator of their estimands based on the EIC. We believe that practitioners that are used to maximum likelihood estimation (MLE) would find it more natural to use a plug-in estimator, and thereby targeted MLE (TMLE). It would be straightforward to develop a TMLE, as the EICs imply how to target initial estimators of  $E(A|L)$  and  $E(Y|A, L)$  for the main terms estimand, and their analogs for the interaction estimand.

## ORCID

Rachael V. Phillips  <https://orcid.org/0000-0002-8474-591X>

Mark J. van der Laan  <https://orcid.org/0000-0003-1432-5511>

## REFERENCES

- Chambaz, A., Neuvial, P. & van der Laan, M.J. (2012) Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6(2012), 1059.
- van der Laan, M.J. (2010) Targeted maximum likelihood based causal inference: part I. *The International Journal of Biostatistics* 6(2), 2.
- Muñoz, I.D. & van der Laan, M.J. (2011) Super learner based conditional density estimation with application to marginal structural models. *The International Journal of Biostatistics* 7(1), 38.
- Rytgaard, H.C., Gerds, T.A. & van der Laan, M.J. (2021) Continuous-time targeted minimum loss-based estimation of intervention-specific mean outcomes. arXiv preprint arXiv:2105.02088.

**How to cite this article:** Phillips, R.V. & van der Laan, M.J. (2022) Rachael V. Phillips and Mark J. van der Laan's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 717–718. Available from: <https://doi.org/10.1111/rssb.12529>

DOI: 10.1111/rssb.12530

# Thomas S. Richardson's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Thomas S. Richardson**

University of Washington, Seattle, Washington, USA

## Correspondence

Thomas S. Richardson, University of Washington, Seattle, Washington, USA.

Email: [thomasr@uw.edu](mailto:thomasr@uw.edu)

I congratulate the authors on an interesting and thought provoking contribution to the literature on modelling associations.

A particular strength of the approach advocated by the authors is that it may facilitate the estimation of main effects on the relative risk scale, by taking  $g = \log$ . In contrast, a GLM specification such as (4) with log link presents difficulties owing to the variation dependence between  $\beta$  and  $\omega(\cdot)$ . This dependence, also called non-congeniality, often leads to computational difficulties in practice (Lumley et al., 2006).

An alternative approach that allows parametric modelling of the log relative risk as a function of base line covariates:

$$\text{RR}(L) \equiv \frac{E[Y | A = 1, L]}{E[Y | A = 0, L]} = \exp(\theta(L; \beta)), \quad (1)$$

is to use the *Odds Product* (OP) as a nuisance model:

$$\text{OP}(L) \equiv \frac{E[Y | A = 1, L]}{(1 - E[Y | A = 1, L])} \frac{E[Y | A = 0, L]}{(1 - E[Y | A = 0, L])} = \omega(L). \quad (2)$$

This has the advantage that the odds product is variation independent of the relative risk; see (Richardson et al., 2017). This specification may also be used as the first stage in a doubly robust estimation method, which permits consistent estimation of the relative risk model given correct specification of the propensity score  $\pi(L)$  and the relative risk model. Though originally formulated for binary exposures, this approach has recently been extended to discrete variables taking finitely many levels and, under monotonicity conditions, continuous exposures taking values in a bounded interval; see (Yin et al., 2021).

While a partially linear GLM using the logit link function does not suffer from the problem of variation dependence it still faces the issue that, owing to the lack of collapsibility of the odds ratio, interpretation of a logistic 'main effect' can be challenging. In particular, even when a specification such as (4) with  $g = \text{logit}$  is correct, the main effect  $\beta$  will never be closer to zero than the log odds ratio for the marginal association.

More generally, when (4) does not hold, the marginal (log) odds ratio may not even lie within the convex hull of the conditional (log) odds ratios as these vary over strata defined by  $L$ . In such a setting, any weighted average of the conditional associations will not reflect the marginal association. Perhaps for this reason, as noted by Lumley et al. (2006), applied researchers often express a preference for relative risks over (log) odds ratios.

## REFERENCES

- Lumley, T., Kronmal, R. & Ma, S. (2006) Relative risk regression in medical research: models, contrasts, estimators, and algorithms. *UW Biostatistics Working Paper Series*. Working paper 293.
- Richardson, T.S., Robins, J.M. & Wang, L. (2017) On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, 519, 1121–1130.
- Yin, J., Markes, S., Richardson, T.S. & Wang, L. (2021) Multiplicative effect modelling: the general case. *Biometrika*, asab064. <https://doi.org/10.1093/biomet/asab064>

**How to cite this article:** Richardson, T.S. (2022) Thomas S. Richardson's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 719–720. Available from: <https://doi.org/10.1111/rssb.12530>

DOI: 10.1111/rssb.12531

# Ilya Shpitser's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Ilya Shpitser**

Johns Hopkins University, Baltimore, USA

## Correspondence

Ilya Shpitser, Johns Hopkins University, Baltimore, USA.

Email: [ilyas@cs.jhu.edu](mailto:ilyas@cs.jhu.edu)

A hallmark of principled causal inference is being careful and explicit about assumptions underlying data analysis. In their timely paper, the authors, Professor Stijn Vansteelandt and Oliver

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](#) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Author. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

Dukes, adopt this view to provide a general roadmap for cautious statistical inference about target parameters. The authors distinguish between assumptions made *for substantive reasons* and those made *for convenience*. An important feature of their approach is that an estimand with respect to an unrestricted model (hereafter, the ‘non-parametric estimand’) must be specified prior to imposing any model restrictions. The authors illustrate their roadmap by defining and estimating main effect and interaction parameters in regression models. These estimands are defined without reference to any model restrictions, but at the same time specialise to known special cases if restrictions hold, for example if the partially linear regression model is true.

The proposal recapitulates developments in mediation analysis, where direct and indirect effect parameters were initially defined as simple functions of regression coefficients in linear causal models, with the definition eventually generalising to a fully non-parametric one based on nested potential outcomes (Pearl, 2001; Robins & Greenland, 1992) or separable effects based on treatment components (Robins & Richardson, 2010).

In causal inference, an estimand in an unrestricted model is often implied by the combination of the causal model, substantive considerations, and identification theory yielding an identified functional for the target causal parameter. The difficulty with the authors’ proposal is that the considerations one should appeal to when defining the non-parametric estimand are unclear. The authors propose a number of desiderata such an estimand must satisfy, but (a) these are perhaps arguable, and (b) given a particular specialised parameter defined in a parametric or semi-parametric model, for instance a regression coefficient, it seems there will be many non-parametric estimands that reduce to the specialised parameter under restricted models, and otherwise satisfy the authors’ desiderata. Different estimands will yield substantively different conclusions, and it is not clear how to choose among them. It would not be surprising if pragmatic considerations (e.g. ease of inference, availability of software) guided such a choice in practice. Indeed, it appears that the authors themselves were guided by pragmatic considerations when choosing among multiple possibilities for the non-parametric interaction parameter. This would seem to be contrary to the spirit of the authors’ proposal.

As a causal inference researcher myself, I applaud the call to clarity about assumptions and estimands made by the authors. However, absent clear substantive guidance or an analogue of identification theory in causal inference, making the crucial choice of one non-parametric estimand out of many appears to be a ‘black art’.

## REFERENCES

- Pearl, J. (2001) Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI-01). San Francisco: Morgan Kaufmann, pp. 411–420.
- Robins, J.M. & Greenland, S. (1992) Identifiability and exchangeability of direct and indirect effects. *Epidemiology*, 3, 143–155.
- Robins, J.M. & Richardson, T.S. (2010) Alternative graphical causal models and the identification of direct effects. *Causality and Psychopathology: Finding the Determinants of Disorders and their Cures*.

**How to cite this article:** Shpitser, I. (2022) Ilya Shpitser’s contribution to the Discussion of ‘Assumption-lean inference for generalised linear model parameters’ by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 720–721. Available from: <https://doi.org/10.1111/rssb.12531>

# Yanbo Tang's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Yanbo Tang**

University of Toronto, Toronto, Ontario, Canada

## Correspondence

Yanbo Tang, University of Toronto, Toronto, Ontario, Canada.

Email: [yanbo.tang@mail.utoronto.ca](mailto:yanbo.tang@mail.utoronto.ca)

The method proposed by Vansteelandt and Dukes is an interesting combination of non-parametric and machine learning techniques with traditional ideas in statistics. We note in this discussion that the authors' argument is easily extended to obtain a quantitative rate in terms of the quality of the nonparametric conditional mean estimate, and comment on how this may impact a practitioner's choice of the machine learning model used to estimate the conditional mean.

It is worth examining the non-parametric estimation problem of the conditional mean more closely, as it reveals the potential trade-off between using complex estimation procedures and the error rate of the normal approximation to the statistic of interest. One of the conditions required in Theorem 2 is

$$\mathbb{P} \left[ \left\{ E(Y|A, L) - \hat{E}(Y|A, L) \right\}^2 \right] = o_p(n^{-1/2}),$$

which (along with the other stated conditions) is sufficient to show that (14) holds, meaning that  $n^{1/2}(\beta - \hat{\beta}) = Z_n + o_p(1)$ , where  $Z_n \xrightarrow{D} N(0, 1)$ . This is crucial for generating confidence intervals with the correct asymptotic coverage.

However the exact rate of consistency of the chosen estimator for  $E(Y|A, L)$  and the other conditional means matters if one wishes to quantify the speed at which the error term decays. For example, consider if all of the conditional means in Theorem 2 have a rate of consistency of  $O_p\{n^{-1/2-\epsilon}\}$ , for some  $\epsilon > 0$ , then following the thread of calculation available in the appendix, we arrive at  $n^{1/2}(\beta - \hat{\beta}) = Z_n + O_p(n^{-\epsilon})$ . While the error in the approximation tends to 0, it may be doing so at a very slow rate which implies a large amount of data is needed to produce a confidence interval with the correct coverage. While if the model were a correctly specified GLM, using the traditional Wald statistic on the regression parameter  $\beta$  would have resulted in the classical error rate of  $O_p(n^{-1/2})$ . This potential loss in the accuracy of the distributional approximation is not surprising as it reflects the difficulty of non-parametric estimation problems in general.

This suggests that the method chosen to estimate the conditional mean needs to be flexible enough to produce a good estimator, while not so complex that the rate of consistency is too slow to be of practical use for a given dataset of size  $n$ . Thus some additional consideration is needed on the practitioner's part when selecting the machine learning or non-parametric method used to estimate the conditional mean, and in particular some knowledge of the true conditional mean function may be required, for example its smoothness in terms of higher-order derivatives.

**How to cite this article:** Tang, Y. (2022) Yanbo Tang's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 722–723. Available from: <https://doi.org/10.1111/rssb.12532>

DOI: 10.1111/rssb.12533

## Eric J. Tchetgen Tchetgen's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Eric J. Tchetgen Tchetgen**

University of Pennsylvania, Philadelphia, Pennsylvania, USA

### Correspondence

Eric J. Tchetgen Tchetgen, University of Pennsylvania, Philadelphia, PA, USA.

Email: [ett@wharton.upenn.edu](mailto:ett@wharton.upenn.edu)

We congratulate Vansteelandt and Dukes (V&D) for an important contribution to the burgeoning literature on methods for valid and interpretable inference about the potential 'effect' of an exposure from a model agnostic perspective. We wholly agree with the general principle of model-free inference about a functional motivated by a semiparametric model, in this case a generalised partially linear model (GPLM). Such a principle is well-grounded and ought to be routinely used in applications of semiparametric theory. In fact, we contend that had this very principle been applied to competing methods considered in the simulation study such as reported in Table 1, a more compelling conclusion might have been reached by clearly separating the near universal advantage of agnostic inference, from the more subjective goal of defining an interpretable non-parametric functional. Specifically, considering the DR estimator of Tchetgen Tchetgen (2013) of section 6.1, we note that a model agnostic (i.e. assumption-lean) inferential framework is readily available upon



deriving the influence function for the non-parametric solution to the population analogue of the DR moment equation. In fact, straightforward algebra reveals that the non-parametric functional implicitly defined by the moment equation is the log of a weighted average of the conditional odds ratio function:

$$\beta = \log \frac{E(w(L)\gamma(L))}{E(w(L))}$$

with weight

$$w(L) = \frac{P(1-P)Q(1-Q)}{c(L)} \geq 0$$

where

$$\gamma(L) = \frac{\Pr(Y = 1 | A = 1, L) \Pr(Y = 0 | A = 0, L)}{\Pr(Y = 1 | A = 0, L) \Pr(Y = 0 | A = 1, L)}$$

is the true conditional odds ratio parameter,

$$\begin{aligned} P &= p(L) = \Pr(Y = 1 | A = 0, L) \\ Q &= q(L) = \Pr(A = 1 | Y = 0, L) \\ c(L) &= \sum_{ay} \gamma(L)^{ay} P^y (1-P)^{1-y} Q^a (1-Q)^{1-a} \gamma(L)^{ay} \end{aligned}$$

Clearly  $\beta = \log \gamma(L)$  only if  $\gamma(L) = \gamma_0$  does not depend on  $L$ , that is under the semi-linear logistic regression model. Agnostic inference about  $\beta$ , can then be obtained by using an empirical version of its efficient influence function (under a non-parametric model for the observed data distribution) possibly combined with cross-fitting as outlined by V&D, which is proportional to

$$\begin{aligned} &\{Y - P\} \{A - Q\} \exp\{-\beta AY\} - E\{\exp\{-\beta AY\} \{Y - P\} | L\} I(Y = 0) \{A - Q\} \\ &- E\{\exp\{-\beta AY\} \{A - Q\} | L\} I(A = 0) \{Y - P\} \end{aligned}$$

Instead of the above influence function, V&D constructed Wald Type CIs omitting the last two terms (which vanish only under the logistic partially linear model), an omission which might explain the poor coverage of DR in scenarios 3 and 4 of Table 1, likely due to excessive bias and under-estimated standard errors.

Finally, it is worth noting that in the case of identity link, Robins et al. (2008), Li et al. (2011) and Mukherjee et al. (2017) proposed minimax estimators of functional (5) of V&D, which are root-n consistent, asymptotically normal and semiparametric efficient under rate conditions for estimating nuisance functions that are significantly weaker than those of V&D in Theorem 2 (nuisance estimation rate conditions for the theorem are therefore sufficient but not necessary at least in the identity link case); their estimator also attains the minimax lower bound of Robins et al. (2009) in non-root-n regimes. More recently, Liu et al. (2021) obtained optimal adaptive minimax estimation results (both in the sense of upper and lower bounds) which simultaneously described both root-n and non-root-n regimes of estimation of a general class of non-parametric functionals which includes (5) in the identity link case. It remains unknown and much of interest whether similar results can be obtained for functional (5) in non-linear link cases.

## REFERENCES

- Li, L., Tchetgen Tchetgen, E.J., van der Vaart, A. & Robins, J.M. (2011) Higher order inference on a treatment effect under low regularity conditions. *Statistics & probability letters*, 81(7), 821–828.
- Liu, L., Mukherjee, R., Robins, J.M. & Tchetgen Tchetgen, E.J. (2021) Adaptive estimation of nonparametric functionals. *Journal of Machine Learning Research*, 22(99), 1–66.
- Mukherjee, R., Newey, W.K. & Robins, J.M. (2017) Semiparametric efficient empirical higher order influence function estimators. arXiv preprint arXiv:1705.07577.
- Robins, J., Li, L., Tchetgen Tchetgen, E.J. & van der Vaart, A. (2008) Higher order influence functions and minimax estimation of nonlinear functionals. In: *Probability and statistics: essays in honor of David A. Freedman*, (pp. 335–421). Durham, NC: Institute of Mathematical Statistics.
- Robins, J., Tchetgen Tchetgen, E.J., Li, L. & van der Vaart, A. (2009) Semiparametric minimax rates. *Electronic journal of statistics*, 3, 1305.
- Tchetgen Tchetgen, E.J. (2013) On a closed-form doubly robust estimator of the adjusted odds ratio for a binary exposure. *American journal of epidemiology*, 177(11), 1314–1316.

**How to cite this article:** Tchetgen Tchetgen, E.J. (2022) Eric J. Tchetgen Tchetgen's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 723–725. Available from: <https://doi.org/10.1111/rssb.12533>

DOI: 10.1111/rssb.12534

# Jiwei Zhao's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Jiwei Zhao**

University of Wisconsin-Madison, Madison, Wisconsin, USA

## Correspondence

Jiwei Zhao, University of Wisconsin-Madison, Madison, WI, USA.

Email: [jiwei.zhao@wisc.edu](mailto:jiwei.zhao@wisc.edu)

First I congratulate Vansteelandt and Dukes for a fascinating and stimulating paper, which I believe will have great impact on the practical data analysis as well as on the application of semi-parametric techniques.

Briefly, the authors propose the following estimand (for main effect)

$$\beta = \frac{E(\text{Cov}[A, g\{E(Y | A, L)\} | L])}{E\{\text{Var}(A | L)\}},$$

and the efficient influence function for  $\beta$  is

$$\frac{\{A - E(A | L)\}[\mu(Y, A, L) - \beta\{A - E(A | L)\}]}{E[\{A - E(A | L)\}^2]}.$$

In the estimation process, other than estimating  $E(Y | A, L)$ , i.e.,  $\hat{E}(Y | A, L)$ , the authors also need the estimate  $\hat{E}(A | L)$ .

In applications, any fitted model could be a misspecified model, for example  $\hat{E}(A | L)$  converges to  $E^*(A | L)$ , where  $E^*$  denotes the expectation that is evaluated under a working model. The similar notations are also used for Cov and Var throughout. Nonetheless, I think the framework the authors propose can be naturally extended to the case where  $E(A | L)$  is misspecified.

We simply generalise the proposed estimand  $\beta$  as

$$\beta^* = \frac{E(\text{Cov}^*[A, g\{E(Y | A, L)\} | L])}{E\{\text{Var}^*(A | L)\}}.$$

Clearly, if the working model  $E^*(A | L)$  is the truth,  $\beta^*$  becomes  $\beta$ . Then, all the properties and theories of  $\beta$  presented in the paper apply. The  $\beta^*$  is motivated from evaluating the mean zero property of the efficient influence function with the misspecified model  $E^*(A | L)$ .

An interesting fact is, when we consider the generalised partially linear model

$$g\{E(Y | A, L)\} = \beta A + \omega(L),$$

then  $\beta^*$  is always equal to  $\beta$ . This means, for the generalised partially linear model, to estimate the estimand of interest, how to model  $E(A | L)$  might not matter too much.

This comment is strongly relevant to the locally efficient estimator in semiparametric models.

**How to cite this article:** Zhao, J. (2022) Jiwei Zhao's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 725–726. Available from: <https://doi.org/10.1111/rssb.12534>

DOI: 10.1111/rssb.12535

# Niwen Zhou and Xu Guo's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

Niwen Zhou<sup>1</sup> | Xu Guo<sup>2</sup>

<sup>1</sup>Center for Statistics and Data Science, Beijing Normal University, Zhuhai, P. R. China

<sup>2</sup>School of Statistics, Beijing Normal University, Beijing, P. R. China

## Correspondence

Xu Guo, School of Statistics, Beijing Normal University, Beijing, P. R. China.

Email: [xustat12@bnu.edu.cn](mailto:xustat12@bnu.edu.cn)

We thank Professors Vansteelandt and Dukes for their innovative and stimulating paper. They introduce a new estimand which reduces to the exposure effect parameter when the (semi)parametric model holds, is generic for continuous or discrete exposure, and still captures conditional association when model assumption fails. We make some additional understanding of the new estimand and introduce a related new estimand.

When there are no covariates  $L$ , a natural estimand is

$$\tau =: \frac{\text{Cov}(A, g\{E(Y | A)\})}{\text{Var}(A)},$$

which reduces to the exposure effect parameter  $\beta$  under model  $g\{E(Y | A)\} = \beta A$ , and is still meaningful as a projection parameter when the model is misspecified (Buja et al., 2019a,b).

When there are covariates  $L$ , consider the conditional version of the above estimand  $\tau$ :

$$\tau(L) =: \frac{\text{Cov}[A, g\{E(Y | A, L)\} | L]}{\text{Var}(A | L)}.$$

To form a meaningful estimand, Vansteelandt and Dukes (2021) considered the ratio of expectations, that is,

$$\frac{E(\text{Cov}[A, g\{E(Y | A, L)\} | L])}{E\{\text{Var}(A | L)\}}.$$

Alternatively, we can consider the following modified estimand

$$\gamma =: E(\tau(L)) = E\left(\frac{\text{Cov}[A, g\{E(Y|A, L)|L\}]}{\text{Var}(A|L)}\right).$$

Here we consider the expectation of the ratio.

Consider a general model:

$$g\{E(Y|A, L)\} = \beta(L)A + \omega(L). \quad (1)$$

Here  $\beta(\cdot)$  and  $\omega(\cdot)$  are unknown functions. Different from the model (4) in Vansteelandt and Dukes (2021), we now allow  $A$ - $L$  interaction, and extend to the exposure effect heterogeneity setting, with respect to  $L$ .

Under the above model (1), the estimand in Vansteelandt and Dukes (2021) is equal to

$$\frac{E\{\beta(L)\text{Var}(A|L)\}}{E\{\text{Var}(A|L)\}}, \quad (2)$$

which is a weighted average of  $\beta(L)$  with weights given by  $\text{Var}(A|L)/E\{\text{Var}(A|L)\}$ , and is generally not equal to  $E[\beta(L)]$ .

While under model (1),  $\gamma = E[\beta(L)]$ , which measures the average exposure effect. When there is no exposure effect heterogeneity, i.e.  $\beta(L) \equiv \beta$ ,  $\gamma$  reduces to  $\beta$  as the estimand in Vansteelandt and Dukes (2021). Even when the model fails, the estimand  $\gamma$  is still meaningful to capture conditional association.

When the exposure is binary,  $g(\cdot)$  is the identity link and moreover  $A$  is independent of the counterfactual outcome  $Y^a$  given  $L$ , our proposed estimand  $\gamma$  reduces to the average treatment effect, i.e.

$$\gamma = E\left(\frac{\pi(L)\{1 - \pi(L)\}(E(Y|A = 1, L) - E(Y|A = 0, L))}{\pi(L)\{1 - \pi(L)\}}\right) = E(Y^1 - Y^0).$$

Clearly, the new estimand in Vansteelandt and Dukes (2021) can be viewed as an extension of weighted average treatment effect; while our new estimand is an extension of the classical average treatment effect.

Lastly, current literature only focus on the conditional expectation of the exposures, which is sensitive to outliers. Conditional quantile can provide a more robust and complete view of the association between response and exposure. Thus it would be very interesting to extend the insightful idea in Vansteelandt and Dukes (2021) to quantile setting.

## REFERENCES

- Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M. et al. (2019a) Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4), 523–544. <https://doi.org/10.1214/18-STS693>
- Buja, A., Brown, L., Kuchibhotla, A.K., Berk, R., George, E. & Zhao, L. (2019b) Models as approximations II: A model-free theory of parametric regression. *Statistical Science*, 34(4), 545–565.
- Vansteelandt, S. & Dukes, O. (2021) Assumption-lean inference for generalised linear model parameters. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*.

**How to cite this article:** Zhou, N. & Guo, X. (2022) Niwen Zhou and Xu Guo's contribution to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 84, 727–729. Available from: <https://doi.org/10.1111/rssb.12535>

The authors replied later, in writing, as follows:

DOI: 10.1111/rssb.12536

## DISCUSSION REPLY

# Authors' reply to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes

**Stijn Vansteelandt** | **Oliver Dukes**

Department of Applied Mathematics, Computer Science and Statistics, Universiteit Gent, Gent, Belgium

### **Correspondence**

Stijn Vansteelandt, Department of Applied Mathematics, Computer Science and Statistics, Universiteit Gent, Krijgslaan 281-S9, Gent 9000, Belgium.

Email: [Stijn.Vansteelandt@ugent.be](mailto:Stijn.Vansteelandt@ugent.be)

We thank all discussants for their interesting and thoughtful comments on our paper. In this rejoinder, we will focus on common themes amongst the commentaries and will close with a discussion of some open issues.

## **1 | TRANSLATING CAUSAL QUESTIONS INTO ESTIMANDS**

Didelez, Shpitser, and Stensrud and Sarvet consider the framework described in our paper as being to some extent at odds with the philosophy of causal inference. There, one translates a

---

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](#) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2022 The Authors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* published by John Wiley & Sons Ltd on behalf of Royal Statistical Society.

scientific question into a causal estimand; ideally this definition should also be model-free. Didelez fears that in spite of the conveniences of our framework, it may lead researchers to bypass the first step of formulating a meaningful question. Stensrud and Sarvet argue that even in simple settings, the parameter that our ‘algorithm’ for generating a target parameter outputs may deviate from the natural causal quantity of interest.

The hygienic causal inference approach is clearly ideal in the simple point-treatment example of Stensrud and Sarvet, and has led to enormous progress in statistical research. In particular, inference for the average treatment effect is generally preferable to our proposal in such settings, provided that ‘treatment’ or ‘no treatment’ are feasible options for all. However, by being somewhat divorced from the specific complexities of the considered data, those hygienic principles can rarely be strictly adhered to; this is especially so as more complex analyses are needed, a point made excellently clear by Daniel. Analyses that were intended to be hygienic, then turn somewhat into a black art, as Shpitser would call it. This is commonly seen in popular marginal structural model analyses for the effect of a time-varying treatment. Here, adjustment for baseline covariates is common, but hardly ever motivated by the strive for a scientifically relevant estimand. It is merely a statistical attempt to trade bias for variance by avoiding the need for inverse weighting to eliminate confounding induced by baseline covariates, which also motivated our work. The lack of guiding principles regarding the choice of baseline covariate set, which Didelez rightly judges to be ad hoc in our proposal, is as much ad hoc in this causal modelling context. Moreover, for similar reasons as explained in Section 2 of our paper, misspecification of the baseline covariate effects in the marginal structural model turns the intended treatment effect estimand into a generally poorly understood functional of the observed data law, which may no longer even summarise that effect. What remains may well be a pale reflection of the hygienic analysis that was intended. We agree with Didelez that problems due to highly variable inverse probability weights elucidate that no useful statements can be made about the considered estimand due to it being too ambitious for the data at hand. However, the sad truth is that these problems often end up being hidden by heuristic truncation of inverse probability weights, which has become the default in software packages. In contrast, our proposal, which could analogously be developed for marginal structural models, does not suffer these problems to the same extent; it is not hiding them, as Didelez seems to worry. This is achieved by targeting an estimand that is not too ambitious for the data at hand. While the weighted average of baseline-covariate-specific treatment effects that we target may appear less appealing, at least it is much better understood than the above marginal structural model estimand when the model is misspecified or inverse probability weights are being truncated.

We agree with Didelez that causal analyses should *ideally* be handled on a case-by-case basis. The difficulties experienced in the above example may in particular be remedied by targeting the effect of specific dynamic interventions designed to be feasible for all. However, translating a question into a causal estimand is often a highly subtle exercise. In practice, even in the causal inference literature, researchers are therefore commonly drawn to estimands that have been well studied, even when their relevance for the scientific question is dubious (e.g. what if the same BMI, or the same level of pack-years of smoking applied uniformly in the population)? It is therefore not uncommon to see exposures being dichotomised/categorised for mathematical convenience, leading to estimands that are deceptively simple, but still remote from the real world (e.g. what if all people in the study population were obese?).

The difficulty of constructing an estimand is further compounded by the fact that some study participants may well be ineligible for the considered interventions, or that the considered exposure cannot be well linked to a specific intervention. The latter is for instance the case in studies

on the effect of obesity. Such studies generally have a causal pursuit, but often of a more exploratory nature; attempting to infer the effect of *specific interventions* on body weight may then go well beyond what the data allow to infer as well as beyond the researchers' initial aim. Our focus on weighted averages of stratum-specific conditional association measures is therefore purposefully less ambitious. The considered weights downweigh individuals for whom few exposure values are plausible. Such individuals would also be less likely recruited if an experiment were conducted. We therefore find Stensrud and Sarvet's example misleading in that there is nothing wrong in finding a treatment effect different from zero when there is treatment effect heterogeneity. In such case, any scalar summary is deficient. Whether the average treatment effect (ATE) is the most relevant target is then context-dependent. It is tempting to believe that conclusions should be drawn for the entire study population and that the marginal causal effect is most relevant (see also Ding), but in practice—even in clinical trials—we often work with convenience samples. Without careful restriction of the study population (Hernán & Robins, 2016), the ATE may well end up focussing on a less relevant population than the retargeted variance-weighted population on which our estimands focus. As such, the considered estimands also address and overcome the formidable task of how to restrict the study population (e.g. consider how difficult it would be to identify individuals in whom obesity could be a plausible exposure status).

Given the difficulty of choosing a proper estimand, we believe that the ideal causal analysis is often not within reach of the many data analysts who have no expert on causal inference within their research network. To connect with applied practice, it is therefore important to provide general purpose strategies that move well beyond the simple binary point-treatment example of Stensrud and Sarvet, while being sufficiently safe to use without necessitating 'black art' remedial measures, such as weight truncation. We believe that our framework offers this. It is partly driven by practical considerations, which Shpitser understood to be against the spirit of our proposal, but whose relevance on the contrary motivated this work. It is a pity that the commentaries did not attempt to demonstrate how an assumption-lean 'algorithm for causal inference', as endorsed by Stensrud and Sarvet, would function in real applied settings that involve more complex queries (e.g. with continuous exposures).

## 2 | CHOICE OF CRITERIA

Several commentators were critical of our three criteria for choosing an estimand. Shpitser suggested that our criteria are 'perhaps arguable' and that in a given problem, there may be many estimands which may satisfy them. We appreciate this concern, but provide clarity. For a given association measure  $\beta(L)$  and weight  $w(L)$  (scaled to have mean 1), we have chosen to focus on weighted averages

$$E\{w(L)\beta(L)\}.$$

These can be interpreted as an average association in a retargeted population that samples individuals with probability proportional to  $w(L)$ . These align well with what we would hope to report—an 'average effect'—when the association  $\beta(L)$  varies with  $L$ , implying that our interpretation of the results would not be grossly misleading if we wrongly assumed  $\beta(L)$  to be constant. Under this model assumption, which appeared to confuse Dong, Gao and Linton, the estimand moreover reduces to a standard model parameter, so that the proposed estimators can also be viewed as root- $n$  consistent estimators in a generalised partially linear model. For the estimand to be more broadly relevant,



we wanted the weight to be the same regardless of the association measure  $\beta(L)$ , so that the same retargeting of the study population applies no matter what outcome is considered. We moreover did not want the weights to depend on features of the outcome distribution, because considerations who to recruit in a study—while often indirectly based on the conditional exposure variability—would not generally be based on the conditional outcome variability. In particular, our choice of  $L$ -specific weights retargets the covariate distribution of the study population to one where all subjects have ‘sufficient’ variation in the exposure. We believe that this retargeted population may well resemble better the population that would be considered in an experiment than the original study population. To evaluate this and to be clear about the population to which the results apply, we recommend reporting summary statistics of the baseline characteristics (e.g. age, gender, etc.) for this retargeted population.

The above criteria, along with the criteria in the paper, leave surprisingly few choices of weights; in fact, we found the construction of an interaction estimand which satisfied all of these criteria to be a non-trivial task. In our proposal, there may however be many ways to define  $\beta(L)$  when  $A$  is not dichotomous. In this paper, we have chosen to work with linear projections as this is visually attractive and drastically simplifies the resulting inference. In future work, we will also consider defining  $\beta(L)$  as the solution to the population maximum likelihood score equation restricted to the stratum  $L$ .

Regarding the first criterion in the paper, Didelez questioned the relevance of choosing an estimand which reduces to a regression coefficient when the model restriction (4) holds. We do not entirely agree. First, there is an abundance of causal queries aimed at developing etiological insight without the immediate ambition of doing a specific intervention. In such settings, it is much easier to reason about the causal data-generating mechanism, than about what estimands might be relevant for the data at hand. This partly explains the popularity of causal diagrams, which enable more intuitive reasoning than that based on counterfactuals. It also explains the popularity of regression-based methods, which continue to dominate applied practice. By connecting to regression models, we believe that we may often connect better to researchers’ a priori understanding of the causal data-generating mechanism, while merely inferring specific features of it. Though the postulated model could be misspecified, our estimands retain close connections to (and sometimes equal) average derivative effects (Hines et al., 2021), which—by virtue of focussing on the effect of a small change in in everyone’s observed exposure—tend to be quite ‘safe’ for general use. An alternative would be to focus on shift interventions that express the effect of increasing the exposure uniformly by, say, 1 unit in the population. The greater appeal of the resulting effects is somewhat deceptive, however, as shift interventions are rarely planned in practice. This then raises the question what would happen when increasing the exposure with 0.5 units, 2 units, ... Flexibly answering these questions calls for some form of modelling, which our framework (when adapted to shift interventions) provides. Second, a key strength of our proposal is that it enables the investigator to work on the scale of choice. In response to Ding’s concern, one may therefore choose to model risk differences or relative risks even for a dichotomous outcome. We agree that the interpretation of model coefficients (causal or otherwise) in more general non-linear models is not always obvious. Nevertheless, if the generalised partially linear model (4) holds, then a given choice of  $\beta$  enables one to work out how specific means or risks  $E(Y|A = 0, L = l)$  in the unexposed would translate into the corresponding means or risks  $E(Y|A = a, L = l)$  at other exposure levels  $a$ . When the model restriction (4) fails, the resulting point estimate may still be useful in terms of giving a rough impression of the strength of association.

For the second criterion, Phillips and van der Laan questioned the importance of choosing estimands for which non-parametric inference does not require estimation of a conditional density. They state that machine learning methods are ‘well-adapted’ for conditional density estimation. Although some proposals have certainly been made, including those by the authors, we are concerned that such estimators may still suffer from unstable performance in finite samples. In fact, unstable performance is often expected with a continuous exposure, even if its conditional density is a priori known, as a result of influential weights for individuals in the tail of the density. Our intention was to develop procedures that could be used safely by non-experts.

For the third criterion, Buja et al. argue that averaging slopes over  $L$ -specific strata is ‘incorrect’. We disagree. Summarising the different slopes obtained for the  $L$ -specific strata in terms of a weighted average is perfectly well aligned with the standard notion of summarising in statistics.

Overall, we agree that other criteria may be worth considering; part of the intention was to stimulate discussion on how to choose an estimand.

### 3 | INTERPRETING THE ESTIMAND

Stensrud and Sarvet, and Phillips and van der Laan argue that the main effect estimand (5) may be difficult to interpret outside of the semiparametric model (4); examples are given where it allegedly fails to capture the causal effect of interest. Stensrud and Sarvet’s example is constructed so that treatment is harmful for half of the population, beneficial for the remainder, and hence the average treatment effect is zero. In contrast, the overlap-weighted treatment effect (6) can be positive or negative depending on whether  $|P(A = 1|L = 1) - 0.5|$  is larger or smaller than  $|P(A = 1|L = 0) - 0.5|$ . Stensrud and Sarvet’s example is designed to illustrate how the proposed estimand may differ ‘from a causal target that more naturally corresponds to (the investigator’s) scientific question of interest.’ However, it is not clear whether either effect is of interest in the presence of *qualitative* effect heterogeneity, particularly when effects are strong. We would argue that conditional/subgroup-specific treatment effects are more useful here. Stensrud and Sarvet’s example highlights the limitations of summary measures, which average (sometimes crudely) over the distribution of  $L$  and/or  $A$ . We believe that there is value in supplementing a summary estimate that provides insight into treatment effect heterogeneity, for example the variance of  $\beta(L)$  in the retargeted population:

$$E[w(L)\{\beta(L) - \bar{\beta}\}^2],$$

for  $\bar{\beta} \equiv E\{w(L)\beta(L)\}$ .

If one is willing to settle for a scalar summary, then it is still questionable whether the average treatment effect best corresponds to the scientific question of interest—at least if the goal is generalisability. Stensrud and Sarvet consider how the overlap-weighted effect changes by changing the conditional distribution of the exposure (given  $L$ ), but fix the distribution of  $L$ . This is reasonable, given the emphasis in causal inference on a well-defined study population. Nevertheless, varying  $P(L = l)$  in their example could also change the direction of the average treatment effect. In cases where the treatment-assignment policy is similar between populations but the covariate distribution changes, as is also reasonable, it is possible that the overlap-weighted treatment effect is better transportable than the average treatment effect.

Phillips and van der Laan also highlight that the numerator of the main effect estimand (5) averages positive and negative contributions  $E(Y|A, L) - E(Y|L)$ , such that the estimand may equal zero in the presence of a strong individual-level treatment effect, and tests of the null hypothesis can suffer from low power relative to tests of other ‘projection-type’ parameters. The example they provide is interesting, but dependent on a lucky choice of reference value (0 in Phillips and van der Laan, and  $x_0$  in Chambaz et al. (2012)). An unlucky choice may likewise make their estimand zero in the presence of a strong individual-level treatment effect. More generally, the connection of our results to the optimality results in Crump et al. (2006) suggest that better power can be expected when the model is correct. In view of the realistic possibility that the model is wrong, we will discuss non-parametric modelling in the next section.

As a brief aside, Phillips and van der Laan also criticise the dependence of the estimand on the conditional distribution of the exposure. However, shift intervention effects, which have been developed in part by those authors (Hubbard & van der Laan, 2008), also share this property. Those estimands consider interventions that transform the conditional distribution of the exposure; instead, we evaluate intervention effects over a retargeted population defined in terms of the conditional distribution of the exposure.

Ding notes that even when the model restriction (4) holds, the estimand (5) will not reduce to a marginal causal effect. It is suggested that the latter parameter is most relevant for policy-making. Our intention was not to wade into the on-going debate about which is most relevant, but we do believe that both marginal and conditional causal effects have advantages and limitations that are important to understand. For example, under model restriction (5), the conditional effect may transport better to different populations since it is insensitive to shifts in the distribution of  $L$ . A weakness of typical approaches for estimating conditional treatment effects is that they rely on parametric modelling assumptions (even in a randomised trial). The imposition of assumptions is understandable, given that these effects are non-pathwise differentiable and therefore the construction of root- $n$  rate non-parametric confidence intervals is not generally feasible. Hence our proposal summarises conditional treatment effects, rendering root- $n$  non-parametric inference feasible.

Hines and Diaz-Ordaz note that our estimands could also be viewed as specific projections of, for example the conditional association between exposure and outcome onto that parameterised by the working model. The concept of projection is indeed highly relevant and has received some attention in the literature on non-parametric inference. In this literature, little or no attention is being paid to the interpretability of the resulting projection estimand. This is especially problematic when, as is commonly done, the entire data-generating model is projected onto the working model. In that case, misspecification in parts of the working model may contaminate all projected model coefficients, as we illustrated in Section 2 of the paper. This is why we have chosen to project merely the conditional association between exposure and outcome (thereby demanding a separate analysis for each considered exposure). The proposal by Hines and Diaz-Ordaz provides a structure for formalising more general estimands along these lines. It will be of interest to understand how specific conditions on the remainder terms in their expansion translate into estimands with specific features.

Responding to Zhou and Guo, we would like to emphasise that our considered estimand explicitly allows for treatment effect heterogeneity by taking a weighted average of conditional treatment effects  $\beta(L)$ . Unlike them, we have chosen not to work with unweighted averages as these do not readily extend to continuous exposures and inference for such effects necessitates inverse probability/density weighting.

## 4 | DATA-ADAPTIVE INFERENCE VERSUS DATA-ADAPTIVE ESTIMANDS

Under model misspecification, Battey, and Lavine and Hodges question whether it is useful to target an estimand for which the interpretation is stable, but which may be misleading about the association of interest. Lavine and Hodges give an example of when the true association between  $Y$  and  $A$  is quadratic; our estimand merely captures the linear association between  $Y$  and  $A$  and so may poorly summarise the data. Battey, and Lavine and Hodges, therefore prefer a sensitivity analysis, which reports the results from multiple models. A key advantage of our proposal is that it avoids the need for such sensitivity analysis with respect to models for the dependence between  $Y$  and the auxiliary covariates  $L$ . However, we are sympathetic towards the concern that a linear (conditional) association between  $Y$  and exposure  $A$  (on the scale of a link function) may sometimes deliver a poor approximation. It is for that reason that the discussion of our paper suggested how the proposal may be extended to estimate that (conditional) association non-parametrically. Even so, the estimation of curves adds complications in view of their high dimensionality, both when it comes to inference and reporting. Our focus on low-dimensional parameters thus remains of interest, even more so as linear approximations are often relevant, for example they sometimes express how much the average outcome would change if each subject's observed exposure were slightly increased (Hines et al., 2021). Sensitivity analyses are useful, but the truth is that subject-matter researchers will often want to present results for a single selected model. In the example of Lavine and Hodges, one may use the data to select a quadratic term in a regression of  $Y$  on  $A$ , but presenting a confidence interval around either the coefficients in the selected model or the model predictions which accounts for the uncertainty in the selection step is then non-trivial; inferential techniques for data-adaptive parameters seem relevant here (Hubbard et al., 2016). Although concepts of sufficiency may be helpful in certain settings, as suggested by Battey, as far as we are aware they cannot be operationalised to account for the many data-adaptive model selection steps that occur in routine data analyses.

## 5 | ESTIMATING NUISANCE PARAMETERS USING MACHINE LEARNING

We appreciate the connection that Hines and Diaz-Ordaz draw to the R-learner, which may potentially aid nuisance parameter estimation.

Bilodeau, Ogburn et al. and Tang claim that the requirement that nuisance parameter estimators converge at a rate faster than  $n^{1/4}$  in our Theorems 2 and 4 may rule out many machine learning methods. These are the same rates discussed elsewhere in the targeted maximum likelihood estimation and debiased machine learning literatures (Chernozhukov et al., 2018; van der Laan & Rose, 2011). Whilst these rates may be attainable in certain contexts (see e.g. Bickel et al. (2009) for sparse estimators, Wager and Walther (2015) for trees and random forests, Chen and White (1999); Farrell et al. (2021) for neural networks), we acknowledge that these results may not reflect how machine learning methods are implemented in practice. Ideally, further developments in statistical learning theory may deepen our understanding of the behaviour of different algorithms in realistic settings. Unfortunately, for applied researchers interested in the implementation of our proposed estimators, it may be overly

challenging to assess the plausibility of the often abstract conditions used in this literature to derive rates.

In the light of this, we give some practical advice. Ideally, a cross-validation-based ensemble method should be used instead of a single candidate learner. Although the simulations in the paper often relied on a single learner, this was done for computational convenience (given the large number of different experiments to run); in the data analysis we were using the Super Learner. Results in van Der Laan and Dudoit (2003), van der Vaart et al. (2006) and van der Laan et al. (2007) suggest that the performance of the Super Learner should be as good as that of the ‘best’ candidate. Like Balzer and Westling (2021), we recommend using a diverse range of candidates, including simple parametric methods and regression splines. These impose greater structure (e.g. linearity or additivity) but may have a faster rate of convergence and better finite sample performance when the requisite assumptions hold. At smaller sample sizes, one may even wish to confine to these simpler methods, which then already improves upon standard analyses by acknowledging post-selection uncertainty. At the moment, we are reassured by observing favourable performance in simulation experiments, but recognise that further, extensive experimentation remains needed.

In challenging, high-dimensional settings, it may be that even the candidate algorithm with the best rate still converges slower than  $n^{-1/4}$ . We therefore agree with Ogburn et al. that an ideal analysis should either supplement inference based on first-order asymptotic theory with sensitivity checks, for example tests of whether bias dominates standard error (Liu et al., 2020), or use alternative approaches that are valid under weaker conditions (Robins et al., 2008). This is an area of exciting development, and further advances will no doubt complement the proposal made here. Of particular interest are methods that are scaleable and can be applied generically by non-experts.

## 6 | COMPARISON WITH ‘PROJECTION’ ESTIMANDS

Battey notes that when the effects of interest are represented by parameters whose interpretations differ according to the model used, the appropriate approach is to acknowledge the model uncertainty rather than seek inference on a quantity whose interpretation is stable but perhaps only tangentially relevant when the assumed model is false. We disagree that the considered parameter is only tangentially relevant. First, it is a weighted average of stratum-specific association measures. Her focus on KL divergence leads to poorly understood estimands, especially in a multivariate sense (see the next paragraph for detail).

Basu and Ding wonder how multivariate parameters would be handled in our proposal. We have purposely chosen to handle one scalar parameter at a time so that a poor projection on one parameter (due to a poor choice of working model) does not contaminate the projections on other parameters. For instance, when the interest lies in the main effect  $\beta$  of  $A$ , a separate analysis is needed from when an interaction  $\gamma$  between  $A$  and some covariate  $Z$  is considered. Moreover, if the interest lies in the sum  $\beta + \gamma$ , then rather than summing the estimates obtained in the two previous analyses, we would derive the efficient influence function of  $\beta + \gamma$  and work with it. This way of working ensures that for instance our inferences for  $\gamma$  do not assume the main effect of  $A$  to be correctly modelled,... This strategy contrasts with typical projection strategies. If simultaneous inferences are nonetheless desired, then inferences can still be developed based on the joint distribution of the efficient influence functions for the different considered estimands.

## 7 | DOUBLE ROBUSTNESS

We appreciate Zhao's suggestion to allow for misspecification of the propensity score, but worry that this is not readily accommodated. The reason is that the derivation of the efficient influence function would require taking directional derivatives of the population limit  $E^*(A|L)$  of the machine learning estimates of the propensity score (under perturbations of the observed data law). Such derivatives would be difficult to obtain as they depend on the features of the considered machine learning algorithm, an issue that we have precisely aimed to avoid.

We find Richardson's alternative parametrisation of the relative risk model attractive compared to the standard approach. However, in our proposal one is free to choose any model/estimator for the nuisance parameters  $E(Y|A, L)$  and  $E[g\{E(Y|A, L)\}|L]$  that may (or may not) respect the constraints on the parameter space. One may fit a logistic model for  $E(Y|A, L)$  and still target a relative risk, for example. This flexibility is important.

Furthermore we are concerned that the resulting inferences and interpretation for the doubly robust estimator developed for the partially linear model in Richardson et al. (2017) are sensitive to violations of model restriction (4). This is especially so if the proposed odds product model is fit data-adaptively, for example using variable selection techniques. For that reason, it may be preferable to seek non-parametric inference for the probability limit of doubly robust estimator, as Tchetgen Tchetgen demonstrates for the odds ratio. Interestingly, when model restriction (4) fails, the resulting odds ratio estimator proposed by Tchetgen Tchetgen no longer appears itself to be consistent if only  $P(Y = 1|A = 0, L)$  or  $P(A = 1|Y = 0, L)$  is consistently estimated. This coheres with our experience that the double robustness properties of semiparametric efficient (or nearly efficient) estimators obtained under the generalised partially linear model may break down outside of the semiparametric model. This includes the 'rate-double robustness' property described, for example by Smucler et al. (2019), where the outcome regression estimator may be allowed to converge at a rate, for example  $n^{1/4}$  or slower, if the propensity score can be estimated at a fast rate (or vice versa). The development of doubly robust methods nevertheless remains useful in our opinion, as we know better how to construct nuisance parameter estimators in this context that target estimators of the parameter of interest with low bias/variance (Cao et al., 2009; Cui & Tchetgen, 2019; Vermeulen & Vansteelandt, 2015).

## 8 | OPEN ISSUES

We agree with Choi and Wong that our data analysis ignored the longitudinal nature of the data. This was so on purpose because we wanted to confine the proposed methodology to generalised linear models. Even so, the extension to longitudinal data models is important and is being worked out along the same principles that we advocate. More generally, we agree with Hennig, Basu, and Zhou and Guo, that extension to more general estimands (e.g. involving quantiles, differences-in-differences) remains needed. With concern for the important problem of influential values, note that the sensitivity of our estimators to such values is readily inspected via histograms of the estimated efficient influence functions.

We are sympathetic to Hunt's remarks. Properly disclosed assumptions may indeed render the analysis honest. Our concern is that model building processes are often complex, and it is generally impossible, even when this is disclosed, to understand how the resulting inference may have been affected. In that sense, we would view the reported confidence intervals as potentially misleading since, even if all assumptions were met and the sample size were large, they would



not cover the truth at the advertised rate. We fully agree that background assumptions, supported by expert knowledge, cannot be avoided in a real data analysis, especially as causal inferences are drawn. Such assumptions are not data-adaptive (i.e. not inferred based on the data being analysed) and were therefore not in the scope of our paper. Finally, we have purposely labelled our methods ‘assumption-lean’ because they do require sufficient smoothness, relative to the size of the data. However, even if parametric methods were considered, we believe that our proposal may still improve upon standard practice by delivering valid post-selection inference when the parametric model holds.

Both the paper and many of the discussions focused on the role of adjustment for confounding. In reality, many data analyses (causal or otherwise) are subject to some form of coarsening (Heitjan & Rubin, 1991), for example missing data, censoring, selection bias, measurement error. In a parametric modelling framework, under a coarsening-at-random assumption we can ignore this bias both in terms of how an estimand is defined, and how inference is done. When the statistical model is incorrect, likelihood-based estimators implicitly target estimands that depend on the coarsening mechanism. These estimands may be inferred with precision, but may be difficult to communicate and compare between studies. In this work, we have deliberately chosen to target estimands that depend on the exposure mechanism, so that they extend to arbitrary exposures. However, in choosing estimands more generally, should we as statisticians prioritise those that are easy to communicate, even if they rely too much on extrapolation? Or should we promote targets that are less ambitious? The answer is not obvious in data subject to more complex coarsening structures, for example if there is non-monotone missingness in  $L$ . Here, non-parametric inference under a ‘missing-at-random’ assumption could be prohibitively complex (Robins, 1997), and a ‘complete case’-type assumption may anyhow be more plausible (Bartlett et al., 2014). Yet such an assumption suggests an estimand that depends on the conditional variance of the exposure given covariates in the complete cases. Borrowing the terminology of Daniel, navigating this ‘bluntness-variance’ trade-off will often be subtle. If one accepts that target parameters should be defined outside of a parametric statistic model, as much of the causal inference community has done, then there is much room for both new estimands and practical guidance in making a choice.

## REFERENCES

- Balzer, L.B. & Westling, T. (2021) Demystifying statistical inference when using machine learning in causal research. *American Journal of Epidemiology*. Available from: <https://academic.oup.com/aje/advance-article/doi/10.1093/aje/kwab200/6322278>
- Bartlett, J.W., Carpenter, J.R., Tilling, K. & Vansteelandt, S. (2014) Improving upon the efficiency of complete case analysis when covariates are MNAR. *Biostatistics*, 15(4), 719–730.
- Bickel, P.J., Ritov, Y. & Tsybakov, A.B. (2009) Simultaneous analysis of Lasso and Dantzig selector. *The Annals of Statistics*, 37(4), 1705–1732.
- Cao, W., Tsiatis, A.A. & Davidian, M. (2009) Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3), 723–734.
- Chambaz, A., Neuvial, P. & van der Laan, M.J. (2012) Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6, 1059–1099.
- Chen, X. & White, H. (1999) Improved rates and asymptotic normality for nonparametric neural network estimators. *IEEE Transactions on Information Theory*, 45(2), 682–691.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., et al. (2018) Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1), C1–C68.
- Crump, R.K., Hotz, V.J., Imbens, G.W. & Mitnik, O.A. (2006) *Moving the goalposts: addressing limited overlap in the estimation of average treatment effects by changing the estimand*. Technical report, National Bureau of Economic Research.
- Cui, Y. & Tchetgen, E.T. (2019) Selective machine learning of doubly robust functionals. *arXiv preprint arXiv:1911.02029*.

- Farrell, M.H., Liang, T. & Misra, S. (2021) Deep neural networks for estimation and inference. *Econometrica*, 89(1), 181–213.
- Heitjan, D.F. & Rubin, D.B. (1991) Ignorability and coarse data. *The Annals of Statistics*, 19(4), 2244–2253.
- Hernán, M.A. & Robins, J.M. (2016) Using big data to emulate a target trial when a randomized trial is not available. *American Journal of Epidemiology*, 183(8), 758–764.
- Hines, O., Diaz-Ordaz, K. & Vansteelandt, S. (2021) Parameterising the effect of a continuous exposure using average derivative effects. *arXiv preprint arXiv:2109.13124*.
- Hubbard, A.E. & van der Laan, M.J. (2008) Population intervention models in causal inference. *Biometrika*, 95(1), 35–47.
- Hubbard, A.E., Kherad-Pajouh, S. & van der Laan, M.J. (2016) Statistical inference for data adaptive target parameters. *The International Journal of Biostatistics*, 12(1), 3–19.
- van der Laan, M.J. & Dudoit, S. (2003) Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: finite sample oracle inequalities and examples. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 130.
- van der Laan, M.J. & Rose, S. (2011) *Targeted learning*. Springer Series in Statistics. New York, NY: Springer New York.
- van der Laan, M.J., Polley, E.C. & Hubbard, A.E. (2007) Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1). Available from: <https://doi.org/10.2202/1544-6115.1309>
- Liu, L., Mukherjee, R. & Robins, J.M. (2020) On nearly assumption-free tests of nominal confidence interval coverage for causal parameters estimated by machine learning. *Statistical Science*, 35(3), 518–539.
- Richardson, T.S., Robins, J.M. & Wang, L. (2017) On modeling and estimation for the relative risk and risk difference. *Journal of the American Statistical Association*, 112(519), 1121–1130.
- Robins, J.M. (1997) Non-response models for the analysis of non-monotone non-ignorable missing data. *Statistics in Medicine*, 16(1), 21–37.
- Robins, J., Li, L., Tchetgen, E. & van der Vaart, A. (2008) Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and statistics: essays in honor of David A. Freedman*, Institute of Mathematical Statistics, pp. 335–421.
- Smucler, E., Rotnitzky, A. & Robins, J.M. (2019) A unifying approach for doubly-robust  $l_1$  regularized estimation of causal contrasts. *arXiv preprint arXiv:1904.03737*.
- van der Vaart, A.W., Dudoit, S. & van der Laan, M.J. (2006) Oracle inequalities for multi-fold cross validation. *Statistics & Decisions*, 24(3), 351–371.
- Vermeulen, K. & Vansteelandt, S. (2015) Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511), 1024–1036.
- Wager, S. & Walther, G. (2015) Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.

**How to cite this article:** Vansteelandt S, Dukes O. Authors' reply to the Discussion of 'Assumption-lean inference for generalised linear model parameters' by Vansteelandt and Dukes. *J R Stat Soc Series B*. 2022;729–739. <https://doi.org/10.1111/rssb.12536>