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

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

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 β 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:

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

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

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 β of A, a separate analysis is needed from when an interaction γ between A and some covariate Z is considered. Moreover, if the interest lies in the sum $\beta + \gamma$, then rather then 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 γ 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