American Journal of Epidemiology

Vol. 173, No. 7
DOI: 10.1093/aje/kwq474
Advance Access publication:
March 16, 2011

Invited Commentary

Invited Commentary: G-Computation–Lost in Translation?

Stijn Vansteelandt* and Niels Keiding

* Correspondence to Dr. Stijn Vansteelandt, Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, S9, 9000 Ghent, Belgium (e-mail: stijn.vansteelandt@ugent.be)

Initially submitted November 3, 2010; accepted for publication November 16, 2010.

In this issue of the *Journal*, Snowden et al. (*Am J Epidemiol*. 2011;173(7):731–738) give a didactic explanation of G-computation as an approach for estimating the causal effect of a point exposure. The authors of the present commentary reinforce the idea that their use of G-computation is equivalent to a particular form of model-based standardization, whereby reference is made to the observed study population, a technique that epidemiologists have been applying for several decades. They comment on the use of standardized versus conditional effect measures and on the relative predominance of the inverse probability-of-treatment weighting approach as opposed to G-computation. They further propose a compromise approach, doubly robust standardization, that combines the benefits of both of these causal inference techniques and is not more difficult to implement.

air pollution; asthma; regression analysis; simulation

Abbreviations: IPTW, inverse probability-of-treatment weighting; SE, standard error.

Standardization in statistics and epidemiology has a long history, going back at least to the 18th century (1). Indirect standardization amounts to applying covariate-specific disease risks for a standard (unexposed) population to the covariate distribution of the study (exposed) population and forms the basic principle underlying the calculation of standardized mortality ratios. The standardized mortality ratio is the ratio between the observed number of cases in the study (exposed) population and the counterfactual number of cases in the study population if it had in fact not been exposed, so that the disease risks in the standard population had applied. Thus, indirect standardization has as its target population the exposed; if the standardized mortality ratio is >1, then the exposure increased disease risk.

Direct standardization goes back to Neison (2) and involves calculating the counterfactual number of cases there would have been in the unexposed population if they had in fact been exposed. By comparing this number with the actual number of cases in the unexposed population, one may again see whether the exposure changes the disease risk, this time with the unexposed as the target population.

Sato and Matsuyama (3) pointed out that the third obvious possibility, using the total population (exposed + unexposed) as the target, corresponds to inverse weighting by the probability of the observed exposure, given the covariates. We emphasize that it also corresponds to the implementation of G-computation in the article by Snowden et al. (4) that, for discrete covariates W and dichotomous (0/1) exposure A and disease status Y, amounts to calculations like

$$\sum_w P(Y=1|A=a,W=w)P(W=w).$$

Here, $P(W=w)$ is estimated as the proportion of subjects with $W=w$ in the sample and $P(Y=1|A=a,W=w)$ as the fitted value from a regression model (the Q-model). As such, the use of G-computation in the article by Snowden et al. (4) is equivalent to what is also known as model-based, smoothed, or regression standardization (5) with the total population as the target.

The above exposition of direct and indirect standardization emphasizes their similarities. However, the following alternative view paves the way for another important distinction in this area, that between conditional and marginal effect measures. Under a multiplicative model for rates,

$$\lambda(t|A=1,W)=\theta\lambda(t|A=0,W),$$

the standardized mortality ratio is the maximum likelihood estimator of θ (6). Clayton (7) suggested from this property that indirect standardization may be viewed as a precursor of confounder control by regression adjustment, for which the effect parameter conditions on the confounders. In contrast, direct standardization and standardization to the total population are used to calculate marginal risks across relevant populations that correspond to crude effect measures from randomized trials. The latter can be seen upon noting that the calculation in the first equation is interpretable as the counterfactual risk of disease, $P(Y_a = 1)$, if all subjects in the relevant population were exposed to $A = a$, provided that—as we will assume from now on—W is sufficient to control for confounding of the effect of A on Y (8).

By applying G-computation, Snowden et al. (4) chose to focus on marginal effect measures. Recall, however, that conditional and marginal effect measures are not always equally conceptually applicable, as Simpson (9; see also reference 10) so entertainingly pointed out. Marginal exposure effects, e.g.,

$$\text{Odds}(Y_1 = 1)/\text{Odds}(Y_0 = 1),$$

are used to evaluate the exposure effect at the level of the study population—as in a randomized trial—and may henceforth be the epidemiologic parameter of interest for making public health policy decisions. However, their interpretation was tied to the particular population across which the marginalization was done. Furthermore, when the exposure is not relevant for a particular subset of the population, it can result in effect measures with limited scientific relevance (11). Conditional exposure effects, e.g.,

$$\text{Odds}(Y_1 = 1|\,W)/\text{Odds}(Y_0 = 1|\,W),$$

are used to evaluate the exposure effect within subpopulations of subjects who share the same covariate value W. They are intended to be more transportable across populations but do require strict assumptions in the case that the regression models are transportable, a "no unmeasured confounder" assumption across all past and future populations to which the results are to be applied.

Current practice in epidemiology has for several decades been dominated by the conditional approach, controlling confounders by stratification, or regression. Conditional and marginal effect measures can nevertheless be quite different, and investigators should carefully consider which to report. The so-called noncollapsibility of certain nonlinear effect measures like the odds ratio implies that conditional effects may differ from marginal effects even in the absence of confounding (12). On the one hand, this may make conditional measures, unlike marginal effect measures, more difficult to compare between studies, because different study results are typically adjusted for different sets of covariates. On the other hand, as previously mentioned, marginal efects may be less transportable between populations. Sizable differences between marginal and conditional effects may also arise as a result of mistakes due to conditioning on intermediate variables, for example, in settings with time-dependent confounding (13, 14), or by inadvertently ignoring covariate exposure interactions (11). In the latter case, estimates of conditional effect measures would roughly weight stratum-specific effects by their precision, whereas marginal effect measures weight stratum-specific effects by the observed covariate distribution in the study population (as in the first equation).

G-COMPUTATION VERSUS INVERSE PROBABILITY-OF-TREATMENT WEIGHTING: COMBINING THE BEST OF BOTH WORLDS?

Historically, G-computation was introduced in a series of revolutionary articles by Robins (13) as a generalization of the first equation to enable adjustment for time-varying confounders that may be intermediate on the causal pathway from earlier exposures to later outcomes. It later formed the theoretical basis of inverse probability-of-treatment weighting (IPTW) in marginal structural models (15). Both G-computation and marginal structural modeling overcome the problem of adjusting for time-varying confounders by avoiding explicit regression adjustment through standardization with the total population as the target. This elucidates that both of these techniques can be used in particular for the standardization of effect measures (3, 5), even when the problem of time-varying confounding is not present. Applications of this nature have nevertheless been rather scarce (see references 3, 16, and 17 for exceptions). The remark in the article by Snowden et al. (4) that G-computation analyses are rare in epidemiology and that the use of IPTW analyses is relatively predominant is thus ambiguous. Although the remark is justified when contemplating the literature on time-varying confounding, for the authors' purpose of standardizing effect measures, IPTW analyses are rare and G-computation is relatively predominant, although commonly termed "standardization."

This division between the use of G-computation and IPTW analysis in these distinct application fields (standardization vs. control for time-varying confounding) has a logical basis. G-computation easily becomes computationally complex as well as demanding in terms of parametric modeling assumptions when time-varying exposures are considered. Further subtleties arise because the outcome regression model or so-called Q-model may be difficult or impossible to match with a marginal structural model in the sense that parsimonious models for the stratum-specific risks, $P(Y = 1|A = a, W = w)$, need not translate into parsimonious models for the standardized risks, $P(Y_a = 1)$ (18). Snowden et al. (4) circumvented these difficulties by restricting their development to binary exposures (and thus to saturated marginal structural models, which are somewhat redundant). In contrast, IPTW analyses for marginal structural models avoid these subtleties by not relying on the Q-model. This makes them more broadly applicable and often the method of choice for the analysis of time-varying exposures. For point exposures, G-computation

and the IPTW approach are viable competitors of similar simplicity. They are equivalent when the covariate W is discrete so that modeling assumptions can be avoided (3), but not otherwise. The IPTW approach is not commonly used in practice because of the traditional reliance on outcome-regression-based analyses, which tend to give more precise estimates. Its main virtue comes when the confounder distribution is very different for the exposed and unexposed subjects (i.e., when there is near violation of the assumption of the experimental treatment assignment), for then the predictions made by the G-computation approach may be prone to extrapolate the association between outcome and confounders from exposed to unexposed subjects, and vice versa. The ensuing extrapolation uncertainty is typically not reflected in confidence intervals for model-based standardized effect measures based on traditional outcome regression models, and thus the IPTW approach may give a more honest reflection of the overall uncertainty (provided that the uncertainty resulting from estimation of the weights is acknowledged) (19). A further advantage of the IPTW approach is that it does not require modeling exposure effect modification by covariates and may thus ensure a valid analysis, even when effect modification is ignored.

In reconciliation of both approaches, we here propose a compromise that combines the benefits of G-computation/model-based standardization and of the IPTW approach. Its implementation is not more difficult than the implementation of these other approaches. As in the IPTW approach, the first step involves fitting a model of the exposure on relevant covariates; this would typically be a logistic regression model. The fitted values from this model express the probability of being exposed and are commonly called ''propensity scores.'' They are used to construct a weight for each subject, which is 1 divided by the propensity score if the subject is exposed and 1 divided by 1 minus the propensity score if the subject is unexposed. The second step involves fitting a model, the Q-model, for the outcome on the exposure and relevant covariates but using the aforementioned weights in the fitting procedure (e.g., using weighted least squares regression). Once estimated, the implementation detailed in the article by Snowden et al. (4) is followed; that is, counterfactual outcomes are predicted for each observation under each exposure regimen by plugging $a = 1$ and then subsequently $a = 0$ into the fitted regression model to obtain predicted counterfactual outcomes. Finally, differences (or ratios) between the average predicted counterfactual outcomes corresponding to different exposure regimens are calculated to arrive at a standardized mean difference (or ratio) (see reference 19 for a similar implementation in the context of attributable fractions).

We refer to this compromise approach as *doubly robust standardization.* Here, the name *doubly robust* expresses that doubly robust standardized effect measures have 2 ways to give the right answer: when either the Q-model or the propensity score model is correctly specified, but not necessarily both. By using a Q-model and at the same time allowing for its misspecification, this approach inherits the benefits of relying on traditional outcome regression models (namely increased precision) but not its limitations (namely the risk of extrapolation bias and the possibility that the Q-model does not match well with the marginal structural model). In a more general context, doubly robust estimators have been criticized for their tendency to amplify model misspecification bias affecting both the Q-model and the propensity score model (21). The doubly robust standardized effect measures proposed here do not cause such bias amplification (19, 22). For the data in the article by Snowden et al. (4), we find that the IPTW approach yields a marginal treatment effect of −0.33 (standard error (SE), 0.062); doubly robust standardization yields similar results: −0.33 (SE, 0.062), −0.34 (SE, 0.062), −0.33 (SE, 0.062), and −0.34 (SE, 0.062), corresponding to regression models 1, 2, 3, and 4, respectively. The fact that the same results were obtained regardless of some of the outcome regression models being misspecified supports the doubly robust nature of this approach.

SUMMARY

In conclusion, the article by Snowden et al. (4) clearly lays out the simplicity of G-computation in the context of point exposures. We hope that their presentation, which is largely divorced from the literature on standardization, will not obscure the equivalence of both techniques. The term *standardization* is revealing and rather well-known to epidemiologists and therefore, in our opinion, is the terminology of choice. The term *G-computation* has so far been mostly reserved to refer to standardization of the effects of time-varying exposures; potentially the term ''G-standardization'' as nomenclature for ''standardization with respect to generalized exposure regimens'' would have been more enlightening. Despite the essential equivalence of G-computation for point exposures and standardization with the total population as the reference, we believe that the developments from the causal inference literature add to the literature on standardization. They give a precise meaning to standardized effect measures in terms of counterfactuals, provide insight into the delicate differences between conditional and marginal epidemiologic effect measures, and suggest novel standardization techniques that combine precision with robustness against model misspecification and extrapolation.

ACKNOWLEDGMENTS

Author affiliations: Department of Applied Mathematics and Computer Science, Ghent University, Ghent, Belgium (Stijn Vansteelandt); Department of Epidemiology and Population Health, London School of Public Health, London, United Kingdom (Stijn Vansteelandt); and Department of Biostatistics, University of Copenhagen, Copenhagen, Denmark (Niels Keiding).

S. V. was supported by Interuniversity Attraction Pole research network grant P06/03 from the Belgian government (Belgian Science Policy).

The authors thank David Clayton for discussions on Simpson's ''paradox'' and on the connections between indirect standardization and regression modeling.

Conflict of interest: none declared.

REFERENCES

1. Keiding N. The method of expected number of deaths, 1786–1886–1986. *Int Stat Rev.* 1987;55(1):1–20.
2. Neison FGP. On a method recently proposed for conducting inquiries into the comparative sanatory condition of various districts, with illustrations, derived from numerous places in Great Britain at the period of the last census. *J Stat Soc (Lond)*. 1844;7:40–68.
3. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. *Epidemiology.* 2003;14(6):680–686.
4. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *Am J Epidemiol*. 2011;173(7):731–738.
5. Rothman KJ, Greenland S, Lash TL. *Modern Epidemiology.* 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008.
6. Kilpatrick SJ. Occupational mortality indices. *Popul Stud*. 1962;16:175–187.
7. Clayton DG. Some remarks on interpretation of models and their parameters in epidemiology. Presented at the XXIst International Biometric Conference, Freiburg, Germany, July 21–26, 2002.
8. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health*. 2006;60(7):578–586.
9. Simpson EH. The interpretation of interaction in contingency tables. *R Stat Soc Ser B*. 1951;13:238–241.
10. Hernán MA, Clayton D, Keiding N. The Simpson's paradox unraveled. *Int J Epidemiol*. In press.
11. Kurth T, Walker AM, Glynn RJ, et al. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am J Epidemiol*. 2006;163(3):262–270.
12. Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. *Stat Sci.* 1999;14(1):29–46.
13. Robins JM. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Math Model*. 1986;7:1393–1512.
14. Cole SR, Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol*. 2002;31(1):163–165.
15. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. *Epidemiology.* 2000;11(5):550–560.
16. Ahern J, Hubbard A, Galea S. Estimating the effects of potential public health interventions on population disease burden: a step-by-step illustration of causal inference methods. *Am J Epidemiol*. 2009;169(9):1140–1147.
17. Taubman SL, Robins JM, Mittleman MA, et al. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009;38(6):1599–1611.
18. Robins JM. Causal inference from complex longitudinal data. Berkane M, ed. *Latent Variable Modeling and Applications to Causality.* (Lecture Notes in Statistics no. 120). New York, NY: Springer Verlag; 1997:69–117.
19. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference [published online ahead of print on November 12, 2010]. *Stat Methods Med Res*. (doi: 10.1177/0962280210387717).
20. Sjölander A, Vansteelandt S. Doubly robust estimation of attributable fractions. *Biostatistics*. 2011;12(1):112–121.
21. Kang JDY, Schafer JL. Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Stat Sci.* 2007;22(4):523–539.
22. Robins JM, Sued M, Lei-Gomez Q, et al. Comment: performance of double-robust estimators when ''inverse probability'' weights are highly variable. *Stat Sci.* 2007;22(4):544–559.