HHS Public Access

Author manuscript

Epidemiology. Author manuscript; available in PMC 2015 November 01.

Published in final edited form as:

Epidemiology. 2014 November ; 25(6): 898–901. doi:10.1097/EDE.0000000000000178.

Applying a Causal Road Map in Settings with Time-dependent Confounding

Maya L. Petersen

Divisions of Biostatistics and Epidemiology, School of Public Health, University of California, Berkeley, CA

Keil and colleagues raise a number of key points in their illustration of the g-formula.[1] In this commentary, I expand on several of these, using a "roadmap" for causal queries that clearly delineates between formal definition of the causal question, translation of this question into a statistical parameter under clearly stated assumptions (identification), estimation of the statistical parameter, and interpretation of the resulting estimate.[2–5] In particular, I discuss approaches to censoring when defining the question, emphasize the longitudinal g-formula as a statistical parameter (identification result) rather than an estimation approach, and discuss targeted maximum likelihood, an alternative approach to estimation of this statistical parameter. I use a simulated example motivated by the data analysis in the paper by Keil et al for illustration and provide accompanying R code in an eAppendix (http://links.lww.com/EDE/A835).

Definition of the question: *What interventions, or changes to the data generating process, are you interested in?*

Keil et al discuss a number of key decisions when defining a causal question, including how best to define the hypothetical interventions of interest. In particular, they make the important point that if the goal is to evaluate the potential impact of a drug that prevents graft-versus-host disease, a hypothetical trial with a control arm in which graft-versus-host disease is allowed to run its natural course provides a more relevant comparison than a control arm in which all patients are somehow forced to develop graft-versus-host disease.

When conceiving of such a hypothetical trial, careful attention must be given not only to interventions on the exposure, but also to interventions on censoring. Consider a trial in which censoring due to loss to follow up is allowed to run its natural course in both the treatment and control arms. Informative loss to follow up can result in biased inferences about the effectiveness of the randomized intervention. For example, perhaps patients whose condition deteriorates over the course of the trial are more likely to drop out, and this effect is particularly pronounced in the control arm. Although the mortality observed in both arms

Correspondence: Maya L. Petersen, Divisions of Biostatistics and Epidemiology, School of Public Health, University of California, 101 Haviland Hall, Ste. 102, Berkeley, Ca 94720. mayaliv@berkeley.edu.

will underestimate the mortality that would have been observed had all patients remained under follow up, the more pronounced underestimate of mortality in the control arm will result in an underestimate of the drug's benefit.

A hypothetical trial as was emulated in the paper by Keil and colleagues, in which loss to follow up was prevented in the treatment arm and allowed to follow its natural course in the control arm, can result in such a differential censoring pattern, potentially obscuring a protective effect of preventing graft-versus-host disease. More generally, such a trial might conclude that a drug with no effect on mortality is either protective or harmful. An additional intervention to prevent censoring in both treatment and control arms would avoid this potential shortcoming. Of course, hypothetical interventions on censoring must also be interpreted carefully, particularly if the censoring process affects the occurrence of the outcome rather than simply whether it is observed.[6,7]

Counterfactuals provide a formal language for expressing causal questions of this nature, including but not limited to questions formulated as hypothetical trials.[8–11] For the remainder of this commentary I focus on counterfactual survival probability at a single time point under an intervention to prevent exposure in all subjects. However, my points also apply to more complex hypothetical interventions, including interventions on censoring, as well as to various summary measures for comparing the distribution of counterfactual outcome under distinct interventions (including the hazard ratio targeted by Keil et al or other marginal structural model parameters).[12]

In particular, I consider a simplified version of the example in the paper by Keil et al, where now the data consist of a continuous platelet level at time 1 (L_1), graft-versus-host disease at time 1 (A_1), death by time 2 (Y_2), platelet level at time 2 (L_2), graft-versus-host disease at time 2 (A_2), and finally death by time 3 (Y_3). We wish to learn the counterfactual probability of having died by time 3 were all graft-versus-host disease prevented at times 1 and 2. In simulated data (such as that provided in the eAppendix, http://links.lww.com/EDE/A835), we can calculate the true value of this quantity by setting the sample size very large, setting $A_1 = A_2 = 0$ in the underlying data generating process and calculating the proportion of individuals surviving, which in this hypothetical example is 7.8% (Figure A).

Identification: *Which statistical parameter equals (or approximates most closely) the causal quantity you care about?*

In contrast to this hypothetical data generating process, graft-versus-host disease occurred in some individuals in the process that generated our observed data, and its occurrence was affected by prior platelet levels (Figure B). Thus even if we had measured an infinite sample of individuals, we could not directly compute the quantity we care about-counterfactual survival in the absence of graft-versus-host disease. We now wish to translate our unobserved counterfactual quantity of interest into a statistical parameter, or, in other words, a parameter of the data generating process we observed. If we succeed, then we have transformed a causal problem into a statistical problem – how best to estimate the true value of this statistical parameter and quantify the uncertainty in our estimates using a finite sample of individuals.

One approach to identifying casual effects is to measure and adjust for confounders (where the randomization assumption or backdoor criteria can be used to formally define whether a given adjustment set is sufficient to control for confounding).[13–15] As reviewed by Keil et al, however, when we care about the joint effect of interventions on multiple exposures, then often no single set of adjustment variables is sufficient to control for confounding of all exposures simultaneously. The transformative insight of the longitudinal G-formula is that in such cases it may still be possible to translate a causal quantity into a statistical parameter by identifying sets of adjustment variables sufficient to control for confounding of each exposure variable in turn.[14,16]

In our example, no single set of confounders satisfies the backdoor criteria for the joint effect of A_1,A_2 on Y_3. For example, L_2 is inadmissible because it is affected by A_1, but if L_2 is omitted from the adjustment set it provides an open back door path (and thus a source of spurious association) between A_2 and Y_3. However, if considered sequentially, L_1 blocks all back door paths from A_1, and (L_1, A_1, L_2, Y_2) blocks all back door paths from A_2. Thus, in our simulation (in fact, in any data-generating process compatible with this directed acyclic graph), the counterfactual probability of death by time 3 under an intervention to prevent graft-versus-host disease is equal to the longitudinal g-formula:

$$\sum_{l_1,l_2,y_2} P(Y_3{=}1|A_1{=}0, A_2{=}0, L_1{=}l_1, L_2{=}l_2, Y_2{=}y_2) \times P(L_2{=}l_2|Y_2{=}y_2, A_1{=}0, L_1{=}l_1) \times P(Y_2{=}y_2|A_1{=}0, L_1{=}l_1) \times P(L_1{=}l_1),$$

(where we use discrete notation and define variables deterministically equal to their last value at time of death).[14,17] The g-formula is a known function of the observed data distribution—if we had a sample of infinite size from the observed data generating process then we could compute the true value of this quantity.

The decision whether to estimate this particular quantity versus a simpler parameter should depend on the scientific question and knowledge of the confounding structure. Is time-dependent confounding plausible and the longitudinal g-formula needed, or would a simpler identification approach suffice? Once the choice of statistical target parameter is made, one can chose among a number of possible estimators of this quantity based on purely statistical considerations such as biasvariance tradeoff and robustness to model misspecification.

Estimation: *Which estimator will estimate the statistical target parameter best?*

Keil and colleagues review one possible approach to estimating the g-formula, often referred to as “parametric g-computation.” This approach relies on estimating each term of the g-formula separately using nonsaturated models and then simulating from these estimates.[18] In realistic settings, simple non-parametric estimation approaches such as saturated regression models are generally not feasible. On the other hand, we rarely know enough to correctly specify lower dimensional parametric models *a priori*, whereas the use of misspecified models can result in bias. Thus, in practice, some process for looking at the data to choose

the “best” regression specification is needed. The challenge increases when, as in our example, an intermediate confounder is continuous and its conditional density rather than conditional expectation must be estimated.

Appropriately wary of bias due to model misspecification, in their worked example, Keil et al choose among a number of regression specifications based on the fit of the survival curve under no intervention (the natural course).[19] As they note, however, a series of misspecified regression models can fit the natural course well while still resulting in a biased effect estimate. A number of alternative approaches are available that automate this process of looking at the data to improve regression specification.[20] In particular, super learning makes it possible to chose among and combine a set of candidate approaches to provide the best fit for each regression in turn.[21]

Importantly, however, neither super learning nor the approach of Keil et al will in general provide a good bias-variance tradeoff for the statistical target parameter itself. Further, no theory supports the validity of the bootstrap (or any other approach) to statistical inference when data-adaptive methods are used to estimate the components of the g-formula (unless, as described below, the initial data-adaptive estimates are updated). The result is a vulnerability of “classic” parametric g computation, as illustrated by Keil et al, to bias and misleading inference.

Happily, alternative approaches to estimating the g-formula are available. In particular, the g-formula can be re-expressed as a series of iterated conditional expectations and thus can be estimated using a series of nested regressions.[22–24] In our example, the g-formula can be rewritten as

$$E_{L_1}\left[E_{L_2,Y_2|A_1=0,L_1}\left[E[Y_3|A_1=0,A_2=0,L_1,L_2,Y_2]|A_1=0,L_1\right]\right].$$

By avoiding the need to estimate conditional densities, this insight alone can reduce bias. However, the fundamental challenge remains: data-adaptive methods are still typically needed to estimate these regressions well, but if used in isolation, they may be overly biased and can undermine the basis for inference.

A targeted maximum likelihood estimator[25] that builds on the double robust iterated conditional expectation estimator of Robins, Bang and Robins[22–24] resolves these challenges by updating the initial regression fits in a step that uses an estimate of the exposure process or propensity score (equivalent to that used with inverse probability weighting). The updating step removes bias, resulting in double robustness – if either the initial sequential regressions or the exposure process are estimated well the estimator will not be meaningfully biased (assuming adequate data support). Updating also allows the sequential regressions to be fit data adaptively while maintaining the basis for valid statistical inferences.

The eAppendix (http://links.lww.com/EDE/A835) provides R code using the ltmle package to apply this targeted maximum likelihood estimator to a simulated sample of size 200.[26,27] In this sample, estimation of the propensity score and the nested regressions using main term

logistic regressions resulted in a point estimate of 8.2% (95% CI of 3.3%, 13.1%), whereas estimation of both using super learning (with candidate approaches including main term logistic models, generalized additive models, Bayesian generalized linear models, k nearest neighbor classification, and forward and backward stepwise selection by AIC) resulted in a point estimate of 8.0% (95% CI 3.3%, 12.8%).[28–30] In this simple data generating process both approaches were unbiased (bias less than 0.001 across 500 repetitions of the simulation); however, in realistic settings there is no guarantee that main term models will perform well and super learning is recommended.

CONCLUSION

Formal casual thinking is most useful when the processes of framing a casual question, translating it into a statistical problem, choosing an estimator, and interpreting results are kept distinct. We have described this process as a "road map."[2,5] Not all roads lead to complex analyses. However, many casual questions correspond to (or are best approximated by) statistical parameters that in realistic sample sizes require complex methods to estimate well. Fortunately, software is increasingly available to facilitate implementation of these methods without extensive coding. Although we have focused on a simple causal parameter, the ltmle package can be used to estimate more complex parameters (including those of marginal structural models) and to implement alternative estimators (including IPW).[26] Of course, software in no way replaces the hard intellectual work required at each step of the process. The roadmap is a guide; used well it can help direct you toward your desired casual destination and understand how far away you may still be.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

I gratefully acknowledge the helpful comments of Mark van der Laan and R code provided by Joshua Schwab.

This work was supported by NIH Grant U01AI069924 (NIAID, NICHD, and NCI) (PIs: Egger and Davies). M.L.P. was supported by a Doris Duke Clinical Scientist Development Award.

References

1. Keil AP, Edwards JK, Richardson DR, Naimi AI, Cole SR. The parametric G-formula for time-to-event data: toward intuition with a worked example. Epidemiology. 2014
2. Petersen ML, van der Laan MJ. Causal models and learning from data: integrating causal modeling and statistical estimation. Epidemiology. 2014; 25:418–426. [PubMed: 24713881]
3. Heckman JJ, Vytlacil EJ. Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. Handb Econom. 2007; 6:4779–874.
4. Pearl, J. The causal foundations of structural equation modeling. In: Hoyle, RH., editor. Handbook of Structural Equation Modeling. New York: Guilford Press; 2012. p. 68-91.
5. van der Laan, M.; Rose, S. Targeted Learning: Causal Inference for Observational and Experimental Data Berlin Heidelberg. New York: Springer; 2011.
6. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15:615–625. [PubMed: 15308962]

7. Hernán MA, Schisterman EF, Hernández-Díaz S. Invited commentary: composite outcomes as an attempt to escape from selection bias and related paradoxes. Am J Epidemiol. 2014; 179:368–370. [PubMed: 24287470]
8. Neyman J. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Stat Sci. 1923 [1990]; 5:465–72.
9. Rubin D. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol. 1974; 66:688–701.
10. Pearl, J. Causality: Models, Reasoning, and Inference. 2. New York: Cambridge University Press; 2000.
11. Hernán MA, Alonso A, Logan R, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008; 19:766–779. [PubMed: 18854702]
12. Robins JM, Hernán MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000; 11:550–560. [PubMed: 10955408]
13. Pearl J. Causal diagrams for empirical research. Biometrika. 1995; 82:669–710.
14. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period. Mathem Mod. 1986; 7:1393–512.
15. Rosenbaum P, Rubin D. The central role of the propensity score in observational studies. Biometrika. 1983; 70:41–55.
16. Robins, J.; Hernan, MA. Estimation of the causal effects of time-varying exposures. In: Fitzmaurice, G.; Davidian, M.; Verbeke, G.; Molenbergh, G., editors. Longitudinal Data Analysis. London, England: Chapman and Hall/CRC; 2009. p. 566-8.
17. Pearl, J.; Robins, J. Uncertainty in Artificial Intelligence 11. San Francisco, CA: Morgan-Kaufman; 1995. Probabilistic evaluation of sequential plans from causal models with hidden variables; p. 444-53.
18. Robins J. Addendum to: A new approach to causal inference in mortality studies with a sustained expsoure period-application to control of the healthy worker survivor effect. Comput Math Appl. 1987; 14:923–945.
19. Taubman SL, Robins JM, Mittleman MA, Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol. 2009; 38:1599–1611. [PubMed: 19389875]
20. Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. New York: Springer-Verlag; 2009.
21. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007; 6 Article25.
22. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics. 2005; 61:962–973. [PubMed: 16401269]
23. Robins J. Commentary on Using inverse weighting and predictive inference to estimate the effects of time-varying treatments on the discrete-time hazard, by Dawson and Lavori. Stat Med. 2002; 21:1663–1680.
24. Robins J. Robust estimation in sequentially ignorable missing data and causal inference models. Proc Am Statist Assoc Sect Bayesian Statist Sci. 2000:6–10.
25. van der Laan MJ, Gruber S. Targeted minimum loss based estimation of causal effects of multiple time point interventions. Int J Biostat. 2012; 8
26. Schwab, J.; Lendle, S.; Petersen, M.; van der Laan, M. ltmle: Longitudinal Targeted Maximum Likelihood Estimation. R Package Version 0.9.3. 2013. Available at: http://cran.r-project.org/web/packages/ltmle/. Accessed 1 July 2014.
27. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; Vienna, Austria: 2013. Available at: http://www.r-project.org/. Accessed 1 July 2014.
28. Hastie, T. gam: Generalized Additive Models. R Package Version 1.09.1. 2013. Available at: http://Cran.r-project.org/package=gam. Accessed 1 July 2014.

29. Gelman, A.; Su, Y-S. arm: Data Analysis Using Regression and Multilevel/Hierarchical Models. R Package Version 1.7–03. 2014. Available at: http://CRAN.R-project.org/package=arm. Accessed 1 July 2014.

30. Venables, WN.; Ripley, BD. Modern Applied Statistics with S. 4. New York: Springer; 2002.

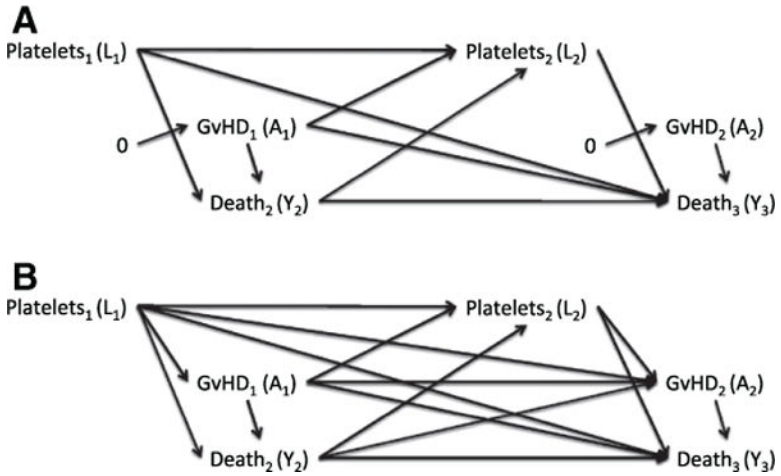

FIGURE.

Directed acyclic graph corresponding to the simulated data example. A, The data generating process under an intervention to prevent graft-versus-host (GvHD) disease at times 1 and 2. B, The observed data-generating process.