Current Epidemiology Reports (2018) 5:205–213
https://doi.org/10.1007/s40471-018-0153-0

EPIDEMIOLOGIC METHODS (R MACLEHOSE, SECTION EDITOR)

A Review of Time Scale Fundamentals in the g-Formula and Insidious Selection Bias

Alexander P. Keil[1] · Jessie K. Edwards[1]

Published online: 15 June 2018

Abstract

Purpose of Review We review recent examples of data analysis with the g-formula, a powerful tool for analyzing longitudinal data and survival analysis. Specifically, we focus on the common choices of time scale and review inferential issues that may arise.

Recent Findings Researchers are increasingly engaged with questions that require time scales subject to left truncation and right censoring. The assumptions necessary for allowing right censoring are well defined in the literature, whereas similar assumptions for left truncation are not well defined. Policy and biologic considerations sometimes dictate that observational data must be analyzed on time scales that are subject to left truncation, such as age.

Summary Further consideration of left truncation is needed, especially when biologic or policy considerations dictate that age is the relevant time scale of interest. Methodologic development is needed to reduce potential for bias when left truncation may occur.

Keywords Causal inference · g-Computation · Time scale · Survival analysis · Longitudinal

Introduction

The introduction of the potential outcome framework has provided a formal set of conditions by which epidemiologists can infer causal relationships from observational data. Using this framework, a number of methods have arisen to estimate causal effects of exposures or treatments in observational data. These methods are often used in longitudinal settings in which exposures, covariates, and outcomes may vary over time. One approach that is fundamental to causal inference is the g-computation algorithm formula (g-formula), which was introduced by Robins, demonstrating that causal inference in complex longitudinal data was possible [1] and which was later formally generalized to estimate effects of arbitrary (e.g., dynamic) treatment regimens [2], competing risks [3], and formal generalization of non-experimental study findings [4]. One of the innovative heuristics that has emerged from this literature has been the re-casting of causal effect estimation in terms of interventions [5]. For epidemiologists, the g-formula has proven to be an essential, if perhaps underutilized tool for estimating human health impacts of exposures and interventions on exposures.

One frequent use for the g-formula is estimating impacts of exposure on the timing of some health event, often referred to as survival analysis. Aside from analytic methods, survival analysis requires three basic tools: a clear definition of failure, a time origin, and a scale for measuring time (time scale) [6]. Such characteristics are often clearly defined in randomized experiments in terms of "natural" time scales like time-on-study, but they may not be in analysis of observational studies with the g-formula where time scale may vary according to substantive interests or for reasons related to policies or interventions of interest; further, multiple time scales may be used as a way to emulate the results of alternative experimental designs [7]. The numerical results and interpretation of a given analysis depend crucially on time scale [8].

In the current manuscript, we explore the choice of time scale in causal effect estimation. Specifically, we discuss issues of time scale for causal effect estimation using the g-formula, illustrating with examples from the literature when possible to demonstrate tradeoffs between etiologic or policy

This article is part of the Topical Collection on *Epidemiologic Methods*

✉ Alexander P. Keil
akeil@unc.edu

[1] Department of Epidemiology, University of North Carolina, 2102E McGavran-Greenberg Hall, Campus Box 7435, Chapel Hill, NC 27599-7435, USA

considerations and the potential for bias. We discuss issues of interpretation that arise when standard complications of survival analysis arise, namely left truncation (late entry) and right censoring (loss to follow-up or end of follow-up). We point out when such issues have arisen in our own and others' work in hopes of spurring stronger consideration of the choice of time scale in causal effect estimation as well as the refinement of methodologic approaches to late entry.

Throughout, we provide examples from a survey of the literature with applications of the g-formula, we performed a MEDLINE search on 9 Oct. 2017 using the search terms given in the Appendix. This search yielded 94 results, which we narrowed to 56 articles after excluding manuscripts that did not include analysis of real data with the g-formula. One of us (AK) classified the articles meeting inclusion criteria according to the time scale of analysis. We focus mainly on the 43 analyses of cohort studies ([9–51], shown in Fig. 1), but we also include selected examples of studies published after the MEDLINE search was conducted. We identified four major types of time scales, given as "study" (time on study), "age," "landmark" (time since some eligibility defining event such as pregnancy), and "treatment" (time since treatment/exposure).

Potential Outcomes and the g-Formula

The g-formula is a powerful tool for causal effect estimation developed by Robins as a way to translate Rubin's potential outcomes framework [52] to the longitudinal setting. Rubin defined potential outcomes with respect to point exposures, or exposures that occur at a specific point in time or represent a summary of exposures measured at a specific point in time. In the point exposure setting, a potential outcome, denoted by y^x, is the value of some outcome (y) we would have observed had exposure been *set* to the value x. For every individual who, in fact, experienced exposure at the level $X = x$, their potential outcome is observed (i.e., everyone has only one observed potential outcome for a given Y, but all other potential outcomes are unobserved, or counterfactual. While causal inference does not require the existence of counterfactuals [53], they are a useful (possible) fiction that greatly facilitates statistical analysis of causal questions [54–56].

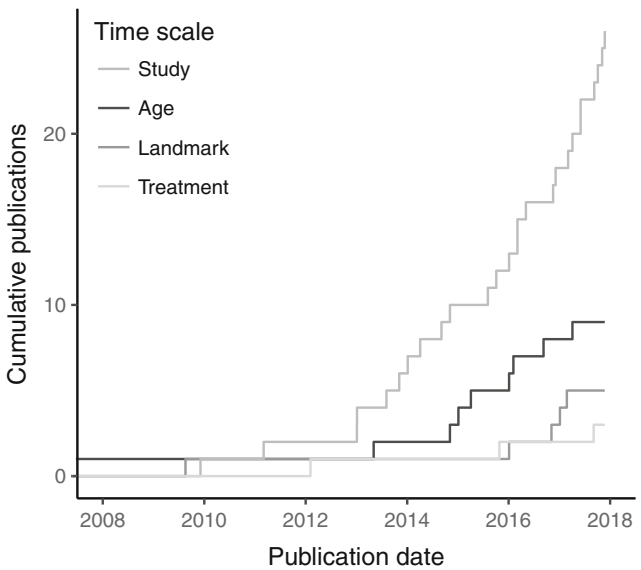

Fig. 1 Cumulative publications of the g-formula identified in a MEDLINE search by time scale of interest. Included studies represent analyses of cohort data (not shown for clarity: reference [9], published in 2001)

In longitudinal settings with time-varying exposures (and confounders), exposure or treatment can be expressed in terms of *regimens*. An exposure regimen can be expressed in terms of a vector $\bar{x}_k = (x_1, \ldots, x_k)$, where the subscript refers to a time index. For example, on the time scale "age," the number of television hours watched at age 10 could be denoted by x_{10}, while the lifetime history of watching television by age 10 could be expressed by $\bar{x}_{10}$. We denote potential outcomes as $\bar{y}_k^{\bar{x}_k}$, which refers to the set of outcomes we would have observed up to time k, had exposure followed the regimen $\bar{x}_k = (x_1, \ldots, x_k)$. Survival potential outcomes, such as time to death, are denoted by $t^{\bar{x}_k}$. For purposes of data analysis with the g-formula, survival outcomes are often discretized such that $\bar{y}_k$ is a vector of binary indicator variables which are 0 except when k is equal to a rounded version of t. This approximation is improved by making the time interval from k to $k+1$ be small relative to t [44]. We denote survival by time k as $\bar{y}_k = 0$.

Assuming no competing risks, the survival function of a potential time to event outcome can be expressed by

$$S^{\bar{x}_k}(k) = \Pr\left(\overline{Y}_k^{\bar{x}_k} = 0\right) = \prod_{m=1}^{k} \Pr\left(Y_m^{\bar{x}_m} = 0 | Y_{m-1}^{\bar{x}_{m-1}} = 0\right) \quad (1)$$

where Π is the product integral and $\Pr\left(Y_0^{\bar{x}_0} = 0\right)$ is defined to be 1. The cumulative distribution function, given as $F^{\bar{x}_k}(k) = 1 - S^{\bar{x}_k}(k)$, is often used for data analysis. In a closed cohort, the cumulative distribution function corresponds to the epidemiologic measure of risk. Implicitly, there is a fixed target population of interest such that the $\overline{Y}_k^{\bar{x}_k}$ refer to potential outcomes in the target population.

The g-formula is often used to estimate risk (sometimes called cumulative risk), demonstrated by the numerous examples, which is useful for inference in public health settings [57•]. The basic unit of risk is given on the right side of (1), $\Pr\left(Y_m^{\bar{x}_m} = 0 | Y_{m-1}^{\bar{x}_{m-1}} = 0\right)$, which is the discrete time hazard at time m under the regimen $\bar{x}_k$. Given data from a closed cohort, the discrete time hazard for potential outcomes under the regime $\bar{x}_m$ can be estimated under the causal identification and

modeling assumptions of the (parametric) g-formula using probability rules and two regression models:

$$\Pr\left(Y_m = 0 \mid \overline{X}_m = \overline{x}_m, \overline{W}_m = \overline{w}_m, Y_{m-1} = 0; \beta\right) \quad (2)$$

$$\Pr\left(W_m = w_m \mid \overline{X}_{m-1} = \overline{x}_{m-1}, \overline{W}_{m-1} = \overline{w}_{m-1}, Y_{m-1} = 0; \alpha\right) \quad (3)$$

where W_m is the set of confounders (assumed to be discrete for convenience) for the effect of X_m on Y_m and β and α are vectors of parameters. Unlike standard approaches to baseline confounding, a model for W_m is necessary in the g-formula. Modeling W_m is typically required to estimate $\Pr\left(Y_m^{\overline{x}_m} = 0 | Y_{m-1}^{\overline{x}_{m-1}} = 0\right)$ because interventions to set $\overline{X}_m$ to $\overline{x}_m$ may impact the distribution of subsequent confounders. Some of the effect of X_m on subsequent outcomes may be through future values of W, and modeling this relationship allows that time-varying confounders may be affected by interventions on exposure. Some analyses target the discrete time hazard ratio, in which case analysis may stop at this point [44, 58].

Using the parametric g-formula to estimate risk involves combining the probability rules of (1) and the parametric models in (2) and (3) to estimate risk under the scenario in which we could set exposure to the regime $\overline{x}_m$. Notably, (2) and (3) can be estimated using standard regression software. The specific model forms used depend on context, but often include familiar model forms such as pooled logistic regression (e.g., [46]), pooled linear regression (e.g., [59, 30]), multi-state models (e.g., [14, 60]), multinomial regression (e.g., [61••]), Cox proportional hazards regression (e.g., [1, 50]), and Bayesian regression (e.g., [62••, 63]). In the case of a point exposure, the g-formula can be simplified to a single model [64]. The point exposure case has become a relatively common example of the g-formula (e.g., [10, 18, 28, 30–32, 34•, 39, 43••, 65–73]); likely because it is a straightforward extension to common modeling frameworks. We focus on the more general case of using the g-formula for survival analysis with time-varying exposure and confounders. However, we cite other examples in which non-survival outcomes (e.g., body mass [26]) are analyzed with the g-formula, but for which issues of time scale are still relevant.

Time and Time Scales in the g-Formula

Time is explicitly considered in the g-formula as the index for time-specific exposures, confounders, and outcomes. The consequence of this is that the value of $\overline{Y}_k^{\overline{x}_k}$ will depend on what the time scale to which k refers. The index k could represent age, calendar period, time on treatment, time on study, time since a landmark event, or any other number of time scales. The origin of any time scale occurs at $k = 0$.

Loss to Follow-up, Censoring, Left Truncation, and Late Entry

In a cohort study, an individual is considered at risk if she or he could (a) experience the outcome of interest and (b) have that outcome recorded in the data. An individual is no longer at risk if he or she experiences the outcome of interest or a competing risk, or if he or she is lost to follow-up. Data are subject to "right censoring" if outcomes might occur among observed individuals but not be recorded. Individuals who are lost to follow-up are often considered censored in that that they are assumed to experience the event of interest after the last time they are observed.

Similarly, individuals can experience outcomes before follow-up starts. Data where some individuals enter the at-risk set after the origin are subject to "left truncation." For non-repeatable survival outcomes, right censoring and left truncation result in outcomes that we do not observe because the individuals experiencing the events are not under observation. Truncation and censoring are distinct, however. Taking mortality outcomes as an example, if an individual has a right-censored death, then he or she may be observed in the study. However, left truncation results in some deaths that occur before an individual could come under observation; then, he or she is prevented from ever being under observation for the study. We take it as a given that the individuals experiencing those events are of interest for estimating causal effects. Thus, when estimating risk, right censoring and left truncation result in under-counting the number of events, while left truncation also results in under-counting the size of the population. Jointly, we refer to left truncation and right censoring as "non-observation." While other types of censoring and truncation may occur, we restrict the remainder of the manuscript to these two processes in order to simplify the exposition.

Non-informative Non-observation and Causal Parameters

Causal inference with the g-formula is possible under a set of sufficient identifying assumptions, namely exchangeability. A simplified version of the exchangeability assumption given by [2] is that

$$\overline{Y}_k^{\overline{x}_k} \coprod (X_k, C_k) | \overline{W}_k = \overline{w}_k, \overline{X}_{k-1} = \overline{x}_{k-1}, C_{k-1} = Y_{k-1} = 0 \quad (4)$$

where $\coprod$ implies independence between the left and the right

sides and C_k is censoring at time k. Notably, no provision for late entry is given in this oft-used definition of exchangeability. However, in his original paper on the g-formula [1], Robins anticipated the problem of late entry, noting that "the only safe option is to match on exposure and [confounder] history until time of entry" (p. 1436). A rigorous derivation of the stronger assumption of exchangeability necessary under late entry is beyond the scope of this paper. Roughly, however, for causal inference to be possible in survival analysis subject to left truncation, exchangeability must hold conditional on the measured past values of exposure and confounders, among those who have entered into the study by time k and remain uncensored at time k. This strengthened exchangeability assumption implies that underlying health status should not depend on entry time, conditional on past values of exposure and confounders. While we do not discuss them here, estimating effects of interventions using the g-formula also requires the causal identification conditions of positivity and consistency and the statistical assumption of correct model specification [1].

From Hazards to Risks with Right Censoring

Provided that exchangeability (4) holds, the conditional, discrete time hazard given in (2) can be consistently estimated given a correct statistical model. However, the survival or cumulative distribution functions may be of greater interest than hazard functions, and we may be more interested in marginal, or population average parameters. Taubman's influential paper on estimating effects of interventions on risk factors for coronary heart disease (e.g., smoking, body mass, and exercise) is a widely cited example that demonstrated how to estimate population average risk using the g-formula [3]. Notably, Taubman et al. used data from the Nurses' Health Study, a cohort study which began follow-up in 1976. The authors used a subset of the person-time from 1982 onwards, but included covariate data from 1980. This method of cohort selection resulted in a cohort that was closed on the left, thus avoiding issues of late entry but allowing for right censoring due to loss to follow-up. Taubman et al. used a version of formula (1), modified to account for competing risks, to estimate the 30-year risk of coronary heart disease in the population under a set of hypothetical interventions that started in 1982.

Taubman's approach to estimating (population average) risk leveraged the absence of late entry. This approach relies on the mathematical result that the population average risk is equal to the average of the individual risks, given by $F^{\bar{x}_k}(k) = n^{-1}\sum_i^n F_i^{\bar{x}_k}(k)$, where i refers to an individual. In the Monte Carlo algorithm often used to estimate parameters of the g-formula, this corresponds to predicting the individual risk at each time k under the intervention "Set $\overline{X}_k = \bar{x}_k$" for all cohort members and then averaging those risks across the population. Naively, this could be problematic in the case of right censoring because censoring results in some participants not being present in the data for some time points (which would result in selection bias from informative censoring due to averaging over potentially increasingly healthy subsets of the population as k grows). Taubman et al. solved this problem by assuming exchangeability and simulating each participants' outcomes beyond the times when they were, in fact, censored (though Taubman et al. also performed a sensitivity analysis in which censoring was not prevented by modeling censoring as a function of prior exposure and covariate histories). Thus, Taubman et al.'s primary analyses encode joint interventions of the type "Set $\overline{X}_k = \bar{x}_k$ and prevent censoring from loss to follow-up." Administrative censoring from the end of follow-up is addressed by simply evaluating risk up to the end of follow-up, presuming it occurs at the same time for all participants on the time scale of interest.

Taubman's approach for loss to follow-up can be easily extended for other forms of censoring, at a possible cost of interpretability. For example, Jain et al. estimate the effect of smoking cessation on 20-year weight gain, which can only be observed among those who survive 20 years [26]. The authors use an implicit intervention to prevent mortality over the risk period. While controversial [74], this approach to competing risks has been shown by Tchetgen Tchetgen et al. to correspond to a combination of principle strata effects and yield a valid test of the sharp null hypothesis between exposure and the outcome of interest [75].

From Hazards to Risks with Left Truncation

For right censoring, the g-formula can account for missing outcomes by simply simulating participant data as though they had not been censored. The situation is not as straightforward for left truncation. Recall that, with right censoring, we are missing outcome data on some individuals who are observed. However, with left truncation, we are missing outcome data on individuals from a target population *who were never observed* because they experienced some event that prevented them from entering the study to begin with (but are nonetheless of interest). For example, in a target population of all workers who were employed at a Montana copper smelter, explored in work by Robins [1] and later by Keil et al. [76], data were available only for workers who were employed on or after 1 January 1938.

Assuming exchangeability, the missing individuals create no problem for estimating the conditional hazard at time k, which is conditional on covariates and on survival to time k and does not depend on the missing data. However, for cumulative average risk, left truncation is more problematic than right censoring because it implies we do not know the size

of the target population to begin with. Note that this problem results when the *target population* is subject to late entry, even when exchangeability holds in the study sample.

Under the definition of exchangeability extended to account for left truncation, Taubman et al.'s algorithm cannot be used to estimate risk in target population data subject to left truncation. Frequently, the study population (reimagined as a closed cohort) is the target population of interest, so the issue may arise in any data subject to left truncation. In order to account for these missing individuals, we would have to simulate extra members of the population, but it is not clear how these pseudo-individuals would be assigned baseline or time-fixed covariates, given that they were never observed. Unfortunately, there is no apparent solution given in the methodologic literature to this exact problem.

To more explicitly define the unique issue caused by late entry, it is helpful to more concretely describe issues with the more tractable problem of right censoring. Consider a possible cause of loss to follow-up in the Nurses' Health Study such as stroke, which could prevent participants from filling out study surveys. Assume that stroke and coronary heart disease (Taubman's outcome of interest) are related through common etiologic pathways, such that they are associated in the population. Under Taubman et al.'s approach, if a woman were lost to follow-up due to stroke-related non-response, her subsequent data would be simulated assuming exchangeability. Thus, in Taubman et al.'s approach, the average target population hazard in older ages would be based on a group of women that includes all of those who survived with stroke. Under an alternative approach, her outcome would be censored after the last completed survey, and risk would be calculated using the hazards estimated among those who were not loss to follow-up due to stroke. If stroke is, in fact, related to underlying CHD risk, then the *average* hazard among older women will have been underestimated due to selection bias caused by censoring those women, even if the conditional hazard ratio is unbiased; Taubman's approach avoids censoring-related selection bias by imputing the covariate values after loss to follow-up, thus eliminating censoring. The alternative approach assumes (with respect to censoring) conditional exchangeability in the study sample, but it assumes *unconditional exchangeability in the target population*. This phenomenon occurs because the target population average risk is calculated from the population average hazards, rather than conditional hazard, and the averaging occurs with respect to the covariate distributions in the observed (and imputed) data. The observed plus imputed data used in Taubman et al.'s approach represent the hypothetical study sample we would have observed, had loss to follow-up been prevented. Notably, this only means that covariates among observed individuals need to be imputed.

The key barrier to progress with reducing selection bias due to late entry in the g-formula is that left truncation means that some members of the target population may never be observed, implying that entire covariate histories of new individuals would need to be imputed. Unfortunately, the theory underlying such an approach is underdeveloped, but addressing potential bias from informative late entry is an area of active research. The analogy with right censoring (and solutions in the g-formula) illuminates this issue and highlights potential solutions. The issues of late entry and exposure occurring prior to baseline are not new, however, and care must be taken when considering both when the time scale starts and when the intervention starts.

Additional Considerations for Interventions in Light of Late Entry

A second problem occurs with the "intervention" aspect of the g-formula with late entry (as well as with causal effect estimation, in general, when exposure occurs before study entry [77]). For the intervention "Set $\overline{X}_k = \overline{x}_k$," we implicitly assume that exposure is modifiable at every time past the origin. Consider, however, an individual who enters an occupational cohort study at time $k = 5$ and, through work records, we discover that individual has been at work, but unexposed from time $k = 0$ through study entry. Also consider the intervention "If at work, be exposed to high levels of silica." Suppose that, if exposed to such levels, this individual would have developed silicosis at time $k = 3$ and left employment. Thus, under the supposed intervention, this individual would not have been in the study sample. Consequently, exposure may influence whether or not an individual enters the study and estimating effects of "intervening" on exposure prior to study entry is subject to selection bias. Such selection bias occurs unless exposure prior to the start of follow-up is known not to affect study outcomes conditional on measured covariates [77], an assumption that is empirically testable if some cohort members do not enter late and exchangeability holds with respect to exposure.

One approach to this issue is to consider exposure accrued prior to study entry to be a baseline confounder, and, as such, cannot be intervened upon. Such issues are common in occupational studies. For example, Neophytou et al. used the g-formula to estimate the effect of interventions on respirable elemental carbon on lung cancer mortality, but only implemented the interventions among the exposure that occurred during observed person-time, while adjusting for exposure that occurred prior to study entry [12]. Such an approach is akin to the "treatment decision design" in which interventions are considered for relevant treatment decisions among prevalent and incident users of drugs [78•]. In other analytic approaches, bias reduction is achieved by creating study samples in which individuals are known to be unexposed/untreated

prior to study entry; this set of approaches has been given the name "new user designs" in observational studies of pharmaceuticals [79]. For some etiologic questions, new user designs reduce bias by eliminating "prevalent" exposure that cannot be intervened upon. However, excluding individuals who enter with prior exposure may not be desirable when dealing with small study samples or when individuals exposed before baseline may represent a scientifically interesting group [80].

Potential for Bias and Time Scale Relevance in the g-Formula

Many examples of cohort data analysis with the g-formula involve questions in which a specific time scale is natural to the study design, reflecting interest in the study population as the implied target population. For example, several authors estimate effects of hypothetical interventions to reduce the population burden of cardiovascular and related diseases (e.g., [20, 29, 35, 36, 42•]). In each of these cases, the proposed interventions are expressed in terms of actions that occur at the beginning of follow-up in a closed (on the left) cohort. For relevance to interventions in the study sample, a natural time scale for analysis is "time-on-study," which was employed by 29/43 of the cohort studies we identified. As shown in Taubman et al.'s example, such analyses can readily address informative loss to follow-up by measured covariates.

In other cases, landmark events such as HIV seroconversion [50], pregnancy [30, 31], or hospitalization [32] define the start of the time scale. In the case of seroconversion and pregnancy, the precise start of the time scale may be unknown, and current approaches to this problem with the g-formula involve imputing the start of the time scale (e.g., [50]). However, in the case of early loss and pregnancy, in particular, such studies have the potential for late entry and possible misidentification of the target population if the time scale might start and stop before measurement of the landmark event [81].

Often, subject matter considerations make it desirable to estimate quantities on time scales that are almost certain to be subjected to late entry in many study designs. Occupational studies are one such example. In the USA, the Occupational Safety and Health Administration (OSHA) sets personal exposure limits for many workplace hazards, such as asbestos [46, 82], radon [23], diesel exhaust [12], arsenic [1, 11], and silica [83]. Exposure limits are based on reducing worker risk below a risk difference of 1/1000 comparing no exposure with exposure at the occupational limit for 45 years (e.g., [84]) where risk is evaluated from ages 20 to 85. Thus, for relevance to policy interventions in the USA, the natural time scale for analysis is "age," even if it does not correspond to many occupational cohort studies that recruit workers based on calendar time; many workers are older than 20 at baseline and some do not even start work by age 20 and so would not be captured. Age was the second most used time scale in the identified analyses (9/43 studies). Of the six studies we identified via MEDLINE with possible late entry [9, 11, 12, 23, 46, 50], five of them utilized age as the time scale of analysis.

Notably, in the occupational studies we identified that employ the g-formula, all [1, 11, 12, 23, 46, 82, 83] but two [14, 22] used age as the time scale of analysis. Consequently, these studies were subject to left truncation because not all workers started employment (and follow-up) at the origin of the time scale, often age 20. Further, workers who were hired at younger ages in each of these studies will be subjected to administrative censoring due to the end of follow-up without reaching the maximum age on the time scale, meaning that not all workers have the same administrative end to the time scale. Typically, such variation in administrative censoring is addressed by considering such censoring to be equivalent to loss to follow-up and requiring that exchangeability must also hold for administrative censoring. Each of these analyses accounted for loss to follow-up using variations of Taubman's approach. All but one of the identified occupational studies with late entry used the implicit assumption that marginal exchangeability held for late entry. The notable exception was the original paper that described the g-formula, which included a data analytic example of estimating lung cancer mortality risk among a cohort of copper smelters exposed to arsenic [1]. In this paper, Robins estimated risk in "the subset of the observed study population hired at age 32 in 1935 remaining on work at high exposure until start of follow-up in 1938" (p. 1503). Thus, in Robins' approach, late entry and variation in administrative censoring times (and the need for additional exchangeability assumptions) were eliminated by considering the target population to be a subset of the study population in which late entry did not occur and who would all reach the end of follow-up on the time scale of interest. This approach trades external validity for internal validity. That is, by conditioning (restricting) on age and date of hire, we lose the ability to easily generalize results to the general smelter worker population. However, through this conditioning, we need only assume that workers in the target population are marginally exchangeable with other workers who could have been in the study (but died or quit work) who were of a similar age, hired at a similar time, and had similar pre-study exposures. Another way to state this idea is that we expect marginal exchangeability to hold within this group of workers who are similar in age, date of hire, and exposure at baseline. An alternative approach might consider completely synthetic worker populations, such as those used by the OSHA to estimate lifetime risk for regulatory purposes.

Conclusions

The choice of time scale in analyses with the g-formula requires careful thought that weighs substantive concerns with the plausibility of the assumptions needed to identify effects on each time scale. Loss to follow-up may not result in bias, provided that exchangeability assumptions are met. However, problems with late entry and variation in administrative censoring times require additional consideration. For the g-formula, in contrast with other survival analysis methods [85, 86], left truncation and right censoring are not automatically addressed by the same method when risk is the estimand of interest and the study population is the target population. We hope this manuscript encourages further consideration of the assumptions necessary to make inference with the g-formula when using a time scale subject to late entry. Further, we hope to spur methodologic innovations to address informative late entry while still allowing inference on biologically or policy relevant time scales.

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Human and Animal Rights and Informed Consent This article does not contain any studies with human or animal subjects performed by any of the authors.

Appendix: MEDLINE search terms

("g-computation" [All Fields] OR "g computation" [All Fields] OR "g-formula" [All Fields] OR "g formula" [All Fields]) AND ("epidemiology" [Subheading] OR "epidemiology" [All Fields] OR "epidemiology" [MeSH Terms]) AND ("English" [Language]).

References

Papers of particular interest, published recently, have been highlighted as:
- • Of importance
- •• Of major Importance

1. 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 Mod. 1986;7(9):1393–512.
2. Young JG, Cain LE, Robins JM, O'Reilly EJ, Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. Stat Biosci. 2011;3(1):119–43.
3. 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(6):1599–611.
4. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. Epidemiology. 2017;28(4):553–61.
5. Hernán MA, Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes. 2008;32(Suppl 3):S8–14.
6. Oakes D. Multiple time scales in survival analysis. Lifetime Data Anal. 1995;1(1):7–18.
7. Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Willett WC, et al. Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology. 2008;19(6):766–79.
8. Westreich D, Cole SR, Tien PC, Chmiel JS, Kingsley L, Funk MJ, et al. Time scale and adjusted survival curves for marginal structural cox models. Am J Epidemiol. 2010;171(6):691–700.
9. Kaufman J, Kaufman S. Assessment of structured socioeconomic effects on health. Epidemiology. 2001;12(2):157–67.
10. Mackey D, Hubbard A, Cawthon P, et al. Usual physical activity and hip fracture in older men: an application of semiparametric methods to observational data. Am J Epidemiol. 2011;173(5):578–86.
11. Keil A, Richardson D. Reassessing the link between airborne arsenic exposure among anaconda copper smelter workers and multiple causes of death using the parametric g-formula. Environ Health Perspect. 2017;125(4):608–14.
12. Neophytou A, Picciotto S, Costello S, et al. Occupational diesel exposure, duration of employment, and lung cancer: an application of the parametric g-formula. Epidemiology. 2016;27(1):21–8.
13. Garcia-Aymerich J, Varraso R, Danaei G, Camargo CA Jr, Hernán MA. Incidence of adult-onset asthma after hypothetical interventions on body mass index and physical activity: an application of the parametric g-formula. Am J Epidemiol. 2014;179(1):20–6.
14. Gran J, Lie S, Øyeflaten I, et al. Causal inference in multi-state models-sickness absence and work for 1145 participants after work rehabilitation. BMC Public Health. 2015;15:1082.
15. Lin S, Young J, Logan R, et al. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. Epidemiology. 2017;28(2):266–74.
16. Schomaker M, Egger M, Ndirangu J, Phiri S, Moultrie H, Technau K, et al. When to start antiretroviral therapy in children aged 2-5 years: a collaborative causal modelling analysis of cohort studies from southern Africa. PLoS Med. 2013;10(11):e1001555.
17. Edwards J, Cole S, Westreich D, Mugavero MJ, Eron JJ, Moore RD, et al. Age at entry into care, timing of antiretroviral therapy initiation, and 10-year mortality among HIV-seropositive adults in the United States. Clin Infect Dis. 2015;61(7):1189–95.
18. Piccolo R, Pearce N, Araujo A, et al. The contribution of biogeographical ancestry and socioeconomic status to racial/ethnic disparities in type 2 diabetes mellitus: results from the Boston Area Community Health Survey. Ann Epidemiol. 2014;24(9):648–54. 54.e1
19. Zhang Y, Young J, Thamer M, et al. Comparing the effectiveness of dynamic treatment strategies using electronic health records: an application of the parametric g-formula to anemia management strategies. Health Serv Res. 2017.
20. Bahia S, Vidal-Diez A, Seshasai S, et al. Cardiovascular risk prevention and all-cause mortality in primary care patients with an abdominal aortic aneurysm. Br J Surg. 2016;103(12):1626–33.
21. Lin S, Young J, Logan R, et al. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. Stat Med. 2017;36(26):4153–66.
22. Norström F, Janlert U, Hammarström A. Is unemployment in young adulthood related to self-rated health later in life? Results from the Northern Swedish cohort. BMC Public Health. 2017;17(1):529.
23. Edwards J, McGrath L, Buckley J, et al. Occupational radon exposure and lung cancer mortality: estimating intervention effects using the parametric g-formula. Epidemiology. 2014;25(6):829–34.

24. Nianogo R, Wang M, Wang A, et al. Projecting the impact of hypothetical early life interventions on adiposity in children living in low-income households. Pediatr Obes. 2017;12(5):398–405.
25. Staerk L, Gerds T, Lip G, et al. Standard and reduced doses of dabigatran, rivaroxaban and apixaban for stroke prevention in atrial fibrillation: a nationwide cohort study. J Intern Med. 2018;283(1): 45–55.
26. Jain P, Danaei G, Robins J, et al. Smoking cessation and long-term weight gain in the Framingham Heart Study: an application of the parametric g-formula for a continuous outcome. Eur J Epidemiol. 2016;31(12):1223–9.
27. Victora C, Horta B, Loret de Mola C, et al. Association between breastfeeding and intelligence, educational attainment, and income at 30 years of age: a prospective birth cohort study from Brazil. Lancet Glob Health. 2015;3(4):e199–205.
28. Dorevitch S, Pratap P, Wroblewski M, Hryhorczuk DO, Li H, Liu LC, et al. Health risks of limited-contact water recreation. Environ Health Perspect. 2012;120(2):192–7.
29. Taubman S, Robins J, Mittleman M, et al. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. Int J Epidemiol. 2009;38(6):1599–611.
30. Westreich D, Cates J, Cohen M, Weber KM, Seidman D, Cropsey K, et al. Smoking, HIV, and risk of pregnancy loss. AIDS. 2017;31(4):553–60.
31. Galin J, Abrams B, Leonard S, et al. Living in violent neighbourhoods is associated with gestational weight gain outside the recommended range. Paediatr Perinat Epidemiol. 2017;31(1):37–46.
32. Jafarzadeh S, Thomas B, Marschall J, et al. Quantifying the improvement in sepsis diagnosis, documentation, and coding: the marginal causal effect of year of hospitalization on sepsis diagnosis. Ann Epidemiol. 2016;26(1):66–70.
33. Lodi S, Sharma S, Lundgren J, Phillips AN, Cole SR, Logan R, et al. The per-protocol effect of immediate versus deferred antiretroviral therapy initiation. AIDS. 2016;30(17):2659–63.
34.• Rogawski E, Meshnick S, Becker-Dreps S, et al. Reduction in diarrhoeal rates through interventions that prevent unnecessary antibiotic exposure early in life in an observational birth cohort. J Epidemiol Community Health. 2016;70(5):500–5. **The authors demonstrate important alternative estimands to "ever/never" contrasts in causal effect estimation. They use the g-formula to estimate the impact of reductions, but not eliminations, of antibiotic exposures in early life on diarrheal illness.**
35. Hubbard A, Jamshidian F, Jewell N. Adjusting for perception and unmasking effects in longitudinal clinical trials. Int J Biostat. 2012;8(2):7.
36. Danaei G, Pan A, Hu F, et al. Hypothetical midlife interventions in women and risk of type 2 diabetes. Epidemiology. 2013;24(1):122–8.
37.• Lodi S, Costagliola D, Sabin C, et al. Effect of immediate initiation of antiretroviral treatment in HIV-positive individuals aged 50 years or older. J Acquir Immune Defic Syndr. 2017;76(3):311–8. **The authors estimated effects of the timing of antiretroviral therapy initiation on a composite clinical outcome. The work demonstrated how observational studies can provide an essential supplement to clinical trials when poor adherence may result in treatment effects that do not reflect real-world benefits.**
38. Lesko C, Todd J, Cole S, et al. Mortality under plausible interventions on antiretroviral treatment and depression in HIV-infected women: an application of the parametric g-formula. Ann Epidemiol. 2017;27(12):783–9 e2.
39. Bielemann R, Gigante D, Horta B. Birth weight, intrauterine growth restriction and nutritional status in childhood in relation to grip strength in adults: from the 1982 Pelotas (Brazil) birth cohort. Nutrition. 2016;32(2):228–35.
40. Lajous M, Willett W, Robins J, et al. Changes in fish consumption in midlife and the risk of coronary heart disease in men and women. Am J Epidemiol. 2013;178(3):382–91.
41. Loret de Mola C, Hartwig F, Gonçalves H, et al. Genomic ancestry and the social pathways leading to major depression in adulthood: the mediating effect of socioeconomic position and discrimination. BMC Psychiatry. 2016;16(1):308.
42.• Danaei G, Robins J, Young J, Hu FB, Manson JE, Hernán MA. Weight loss and coronary heart disease: sensitivity analysis for unmeasured confounding by undiagnosed disease. Epidemiology. 2016;27(2):302–10. **The authors demonstrate a simple sensitivity analysis for unmeasured confounding using the g-formula. This approach represents a useful way to deal with problems that arise due to undiagnosed disease and frailty.**
43.•• De Stavola B, Daniel R, Ploubidis G, et al. Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. Am J Epidemiol. 2015;181(1): 64–80. **The authors relate a classical method for causal inference, structural equation modeling, to the parametric g-formula. This work contrasts the benefits and drawbacks of two of the primary schools of causal effect estimation, a barrier to progress in the applied causal effect estimation literature. The authors demonstrate both methods in a mediation example.**
44. Keil A, Edwards J, Richardson D, et al. The parametric g-formula for time-to-event data: intuition and a worked example. Epidemiology. 2014;25(6):889–97.
45.• Liu W, Zhang Z, Schroeder R, et al. Joint estimation of treatment and placebo effects in clinical trials with longitudinal blinding assessments. J Am Stat Assoc 2016;111(514):538–48. **The authors demonstrate an approach to quantifying effects in clinical trials that are due to the structure of the clinical trial itself. This work presents an important bridge between randomized experiments and observational studies, which may be important for formal data fusion between these two lines of evidence.**
46. Cole S, Richardson D, Chu H, et al. Analysis of occupational asbestos exposure and lung cancer mortality using the g formula. Am J Epidemiol. 2013;177(9):989–96.
47. Schomaker M, Leroy V, Wolfs T, Technau KG, Renner L, Judd A, et al. Optimal timing of antiretroviral treatment initiation in HIV-positive children and adolescents: a multiregional analysis from Southern Africa, West Africa and Europe. Int J Epidemiol. 2017;46(2):453–65.
48. Westreich D. From exposures to population interventions: pregnancy and response to HIV therapy. Am J Epidemiol. 2014;179(7): 797–806.
49. Schomaker M, Davies M, Malateste K, Renner L, Sawry S, N'Gbeche S, et al. Growth and mortality outcomes for different antiretroviral therapy initiation criteria in children ages 1-5 years: a causal modeling analysis. Epidemiology. 2016;27(2):237–46.
50. van der Wal W, Prins M, Lumbreras B, Geskus RB. A simple G-computation algorithm to quantify the causal effect of a secondary illness on the progression of a chronic disease. Stat Med. 2009;28(18):2325–37.
51. Lodi S, Phillips A, Logan R, Olson A, Costagliola D, Abgrall S, et al. Comparative effectiveness of immediate antiretroviral therapy versus CD4-based initiation in HIV-positive individuals in high-income countries: observational cohort study. Lancet HIV. 2015;2(8):e335–43.
52. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66(5):688–701.
53. Dawid AP. Causal inference without counterfactuals. J Am Stat Assoc. 2000;95(450):407–24.
54. Rubin DB. Causal inference without counterfactuals: comment. J Am Stat Assoc. 2000:435–8.
55. Robins JM, Greenland S. Causal inference without counterfactuals: comment. J Am Stat Assoc. 2000;95(450):431–5.
56. Pearl J. Causal inference without counterfactuals: comment. J Am Stat Assoc. 2000:428–31.

57.• Cole SR, Hudgens MG, Brookhart MA, Westreich D. Risk. Am J Epidemiol. 2015;181(4):246–50. **The authors describe risk as an epidemiologic measure. They describe ways to estimate risk and make the case that risk is fundamental to understanding how exposures influence the transition from health to disease.**

58. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2012;31(18):2000–9.

59. Horta B, Schaan B, Bielemann R, et al. Objectively measured physical activity and sedentary-time are associated with arterial stiffness in Brazilian young adults. Atherosclerosis. 2015;243(1):148–54.

60. Chen H, Gao S. Estimation of average treatment effect with incompletely observed longitudinal data: application to a smoking cessation study. Stat Med. 2009;28(19):2451–72.

61.•• Edwards JK, Cole SR, Moore RD, Mathews WC, Kitahata M, Eron JJ Sensitivity analyses for misclassification of cause of death in the parametric g-formula. Am J Epidemiol. 2018. **The authors present a new way of modeling outcomes in the framework of the g-formula. They use this new approach to demonstrate a sensitivity analysis for outcome misclassification. This work presents an important tool for performing causal effect estimation with imperfect data.**

62.•• Keil AP, Daza EJ, Engel SM, Buckley JP, Edwards JK A Bayesian approach to the g-formula. Stat Methods Med Res. 2017. **The authors demonstrate an algorithm for estimating the parameters of the g-formula in a fully Bayesian framework. They demonstrate the advantages of this approach in simulations of correlated exposures and small, longitudinal datasets. The authors demonstrate that a number of existing Bayesian hierarchical methods can be used within a causal effect estimation framework.**

63. Wang W, Scharfstein D, Wang C, Daniels M, Needham D, Brower R, et al. Estimating the causal effect of low tidal volume ventilation on survival in patients with acute lung injury. J R Stat Soc Ser C Appl Stat. 2011;60(4):475–96.

64. 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):731–8.

65. Treves-Kagan S, El A, AM PA, et al. Gender, HIV testing and stigma: the association of HIV testing behaviors and community-level and individual-level stigma in rural South Africa differ for men and women. AIDS Behav. 2017;21(9):2579–88.

66. Zhang Y, Laraia B, Mujahid M, et al. Does food vendor density mediate the association between neighborhood deprivation and BMI?: a G-computation mediation analysis. Epidemiology. 2015;26(3):344–52.

67. Patel M, Westreich D, Yotebieng M, et al. The impact of implementation fidelity on mortality under a CD4-stratified timing strategy for antiretroviral therapy in patients with tuberculosis. Am J Epidemiol. 2015;181(9):714–22.

68. Leslie H, Karasek D, Harris L, et al. Cervical cancer precursors and hormonal contraceptive use in HIV-positive women: application of a causal model and semi-parametric estimation methods. PLoS One. 2014;9(6):e101090.

69. Austin P, Urbach D. Using G-computation to estimate the effect of regionalization of surgical services on the absolute reduction in the occurrence of adverse patient outcomes. Med Care. 2013;51(9):797–805.

70. Brewer N, Zugna D, Daniel R, Borman B, Pearce N, Richiardi L. Which factors account for the ethnic inequalities in stage at diagnosis and cervical cancer survival in New Zealand? Cancer Epidemiol. 2012;36(4):e251–7.

71. Wang A, Nianogo R, Arah O. G-computation of average treatment effects on the treated and the untreated. BMC Med Res Methodol. 2017;17(1):3.

72. Fleischer N, Fernald L, Hubbard A. Estimating the potential impacts of intervention from observational data: methods for estimating causal attributable risk in a cross-sectional analysis of depressive symptoms in Latin America. J Epidemiol Community Health. 2010;64(1):16–21.

73. Wang A, Arah O. G-computation demonstration in causal mediation analysis. Eur J Epidemiol. 2015;30(10):1119–27.

74. Chaix B, Evans D, Merlo J, Suzuki E. Commentary: Weighing up the dead and missing: reflections on inverse-probability weighting and principal stratification to address truncation by death. Epidemiology. 2012;23(1):129–31.

75. Tchetgen EJT, Glymour MM, Shpitser I, et al. Rejoinder: to weight or not to weight?: on the relation between inverse-probability weighting and principal stratification for truncation by death. Epidemiology. 2012;23(1):132–7.

76. Keil AP, Richardson DB. Reassessing the link between airborne arsenic exposure among anaconda copper smelter workers and multiple causes of death using the parametric g-formula. Environ Health Perspect. 2016;125(4):608–14.

77. Flanders WD, Klein M. Properties of 2 counterfactual effect definitions of a point exposure. Epidemiology. 2007;18(4):453–60.

78.• Brookhart MA. Counterpoint: the treatment decision design. Am J Epidemiol. 2015;182(10):840–5. **The author proposes a pharmacoepidemiologic study design that leverages data from prevalent users and re-casts causal questions by reconsidering the time scale of interest and the types of interventions of interest.**

79. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003;158(9):915–20.

80. Vandenbroucke J, Pearce N. Point: incident exposures, prevalent exposures, and causal inference: does limiting studies to persons who are followed from first exposure onward damage epidemiology? Am J Epidemiol. 2015;182(10):826–33.

81. Kinlaw AC, Buckley JP, Engel SM, Poole C, Brookhart MA, Keil AP. Left truncation bias to explain the protective effect of smoking on preeclampsia: potential, but how plausible? Epidemiology. 2017;28(3):428–34.

82. Richardson DB, Keil AP, Cole SR, Dement J. Asbestos standards: impact of currently uncounted chrysotile asbestos fibers on lifetime lung cancer risk. Am J Ind Med. 2018;61:383–90.

83. Keil AP, Richardson DB, Westreich D, Steenland K. Estimating the impact of changes to occupational standards for silica exposure on lung cancer mortality. Epidemiology. 2018.

84. Occupational Safety and Health Administration. Occupational exposure to respirable crystalline silica. Final rule. In: Occupational Safety and Health Administration, ed. Fed Regist 2016;81(58).

85. Kaplan E, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958;53(282):457–81.

86. Therneau TM, Grambsch PM. Modeling survival data: extending the Cox model. New York: Springer; 2000.