Title:

The parametric G-formula for time-to-event data: towards intuition with a worked example

Authors:

Alexander P. Keil[a], Jessie K. Edwards[a], David R. Richardson[a], Ashley I. Naimi[b], Stephen R. Cole[a]

Affiliations:

[a]Department of Epidemiology, University of North Carolina, Chapel Hill, North Carolina; Chapel Hill, North Carolina and [b]Department of Epidemiology, Biostatistics, and Occupational Health, McGill University; Montréal, Québec

Corresponding Author:

Alexander Keil, Department of Epidemiology, CB 7435, University of North Carolina, Chapel Hill, NC 27599-7435. E-mail: akeil@unc.edu. T: 919-966-6652. F: 919-966-2089

Running head:

G-formula: intuition and worked example

Sources of financial support:

This work was supported partially by the National Institute of Health Sciences (Grant #s T32 ES007018, R01-CA117841, R01-AI100654) and Post-Doctoral Research Award from the Fonds de recherche en santé du Québec

Acknowledgements:

Igor Karp for helpful comments on a presentation of results from the manuscript.

ABSTRACT

Background: The parametric g-formula can be used to estimate the effect of a policy, intervention, or treatment. Unlike standard regression approaches, the parametric g-formula can be used to adjust for time-varying confounders that are affected by prior exposures. To date, there are few published examples in which the method has been applied.

Methods: We provide a simple introduction to the parametric g-formula and illustrate its application in analysis of a small cohort study of bone marrow transplant patients in which the effect of treatment on mortality is subject to time-varying confounding.

Results: Standard regression adjustment yields a biased estimate of the effect of treatment on mortality relative to the estimate obtained by the g-formula.

Conclusions: The g-formula allows estimation of a relevant parameter for public health officials: the change in the hazard of mortality under a hypothetical intervention, such as reduction of exposure to a harmful agent or introduction of a beneficial new treatment. We present a simple approach to implement the parametric g-formula that is sufficiently general to allow easy adaptation to many settings of public health relevance.

INTRODUCTION

Imagine an oncologist knocks on your door with the following problem: she wants to know how much she could reduce mortality among her bone marrow transplant patients by prescribing a new drug that prevents graft-versus-host disease, a side effect of allogeneic marrow transplantation.[1] While graft-versus-host disease is associated in observational studies with an increased risk of mortality, it also reduces the risk of leukemia relapse – thus, any drug that prevents graft-versus-host disease may have the very undesirable side effect of increasing the rate of relapse.[2] She wants to compare the mortality in her cohort with what mortality would be in that same cohort if they had taken this new drug. We cannot answer this question with a regression model because leukemia relapse is a risk factor for mortality and subsequent graft-versus host disease and it will also decrease the incidence of subsequent relapse (i.e. relapse is a confounder affected by exposure).[3, 4] However, we can answer this question using the g-formula.

The g-formula is an analytic tool for estimating standardized outcome distributions using covariate (exposure and confounders) specific estimates of the outcome distribution.[5]The g-formula can be used to estimate familiar measures of association, such as the hazard ratio. In the current paper, we address the oncologist's question: we compare observed mortality in our cohort with the expected mortality in that cohort under the new treatment.

Epidemiologists often use regression models (for example, the Cox proportional hazards model) to adjust for confounding, equivalent to estimating stratum-specific hazard ratios and then averaging the information-weighted hazard ratios. When some of those confounders are also causal intermediates, this amounts to adjusting away some of the effect of exposure.[6, 7] The g-formula works differently: first, one finds weighted averages of the stratum specific hazards, and then those averaged (standardized) hazards are combined in a summary hazard ratio. Thus, bias resulting from time-varying covariates that can be both confounders and causal intermediates is a shortcoming of using regression models to control for confounding, rather than a general principle of observational data analysis.[8, 9] The g-formula is a tool that overcomes this shortcoming, but its use in the literature has been sparse – we could only find 9 examples using observational data.[8, 10-17] We hypothesize that the dearth of software packages and lack of useful, yet simple examples of the g-formula have been the main barriers to broader use.

We show how the g-formula can be used with standard software tools that many epidemiologists already employ, and we illustrate it using publicly-available data from a small cohort study with accompanying SAS code in an eAppendix. We illustrate how we can estimate the net (total) effect of a hypothetical treatment to prevent graft-versus-host disease on mortality and compare the g-formula approach with a regression approach. The g-formula (as with any statistical method) relies on making assumptions in order to make sense of the complex processes underlying the data. We discuss possible ways to assess how well we meet the assumptions as well as the robustness of the g-formula to violations of these assumptions.

METHODS

The g-formula

Using regression methods to control confounding requires making the assumption that the effect measure is constant across levels of confounders included in the model. Alternatively, standardization allows us to obtain an unconfounded summary effect measure without requiring this assumption. The g-formula is a generalization of standardization and can be expressed similarly. For example, the 10-year risk of death for a group of individuals, standardized across some dichotomous (1,0) risk factor Z could be expressed as

$$\Pr(death = 1) = \sum_{z} \Pr(death = 1|Z = z)\Pr(Z = z)$$

where Σ_z indicates that we are summing over each possible value of Z, and $\Pr(Z = z)$ is the probability that Z takes on the value z in the reference population. If the 10-year risk of death among the group with Z=1 was 0.1 and the risk among the group with Z=0 was 0.05, and we were studying a population with a 60 individuals with Z=1 and 40 individuals with Z=0, the Z-standardized 10-year risk of death would be 0.1*(60/100)+0.05*(40/100) = 0.08.

The g-formula relies on this same set of calculations. Suppose now that we have a cohort of 50 year old men and we wish to estimate the change in 10-year risk of death from some exposure, *X*, that has no effect on mortality among 50-55 year olds but increases mortality after age 55. The overall 10-year risk of death, Pr($death_{10}$=1), could be calculated using the probabilities expressed by the g-formula:

$$\Pr(death_{10} = 1) = \sum_{x} \Pr(death_{10} = 1|X = x, age > 55, death_5 = 0)\Pr(X = x|age > 55, death_5 = 0)$$
$$\times (1 - \Pr(death_5 = 1|X = x, age \leq 55)\Pr(X = x|\ age \leq 55))$$
$$+\ \Pr(death_5 = 1|X = x, age \leq 55)\Pr(X = x|\ age \leq 55)$$

Assume 50% are exposed at the start. As shown in the calculations from Table 1, if the probability of death in our cohort after 5 years (at age 55 - $death_5$) is 0.025 for everyone and the conditional probability of death between 5 and 10 (i.e. between age 55 and 60) years of follow-up is 0.06 for exposed and 0.051 for unexposed individuals, the overall standardized 10-year risk of death ($\Pr(death_{10} = 1)$) would be 0.08. We consider the probability of death in the second period to be conditional because we condition (restrict) our data to the individuals still alive at the beginning of year 6. The overall standardized risk is often referred to as the “natural course”, because we are estimating the risk under no interventions. In this setting (in which we can observe all covariate specific probabilities), the observed 10-year risk will equal the 10-year risk calculated using the “natural course”.

Use of the g-formula for effect estimation proceeds using the standardization formula above and substituting new values for probabilities of exposure defined by a hypothetical intervention. For example, if we wished to estimate the effect of exposure on the 10-year risk of death, we would use the g-formula to calculate the risk of death if we intervened to make all individuals “always exposed “ (i.e. set $\Pr(X = 1|age) = 1.0$ (and $\Pr(X = 0|age) = 0.0$) regardless of age) versus the risk if we intervened to make all individuals “never exposed” (e.g. set $\Pr(X = 0|age) = 1.0$ (and $\Pr(X = 1|age) = 0.0$). As shown in Table 1, we substitute 1.0s (and 0.0s) into the standardization formula for these conditional exposure probabilities and arrive at a 10-year risk for the intervention “always exposed” of 0.084 and for the intervention “always unexposed” of 0.075, (g-formula standardized risk difference = 0.009). For an intervention in which we expose all participants in the last 5 years only (the treatment plan “expose if still alive at 5 years”), we would “intervene” by setting conditional probabilities $\Pr(X|age > 50)$ to 1.0 and $\Pr(X|\ age \leq 55)$ to the conditional proportions observed in the original data (risk=0.084).

The parametric g-formula

While standardization using the g-formula is simple enough to do by hand in our hypothetical example, epidemiologic studies often have a richer covariate sets and longer periods of follow-up. In such studies, stratification by many covariates can quickly lead to strata in which there are not enough data to calculate the conditional probabilities, and use of continuous covariates does not allow this stratification approach. Instead of directly calculating every conditional probability, we can use parametric regression modeling to calculate the conditional probabilities used to carry out computations shown above. Further, we can simulate data from the conditional probability distributions to approximate standardization. Robins referred to this approach of using modeled conditional probabilities to estimate standardized effect measures as the parametric g-formula.[14, 18] The data from our illustrative example and SAS code to carry out our algorithm are provided in eAppendices 1 through 4.

Illustrative example

We use the parametric g-formula to estimate the hazard ratio comparing mortality in a cohort of bone marrow transplant recipients (described in detail by Copelan and Klein and Moeschberger) under different treatments by a hypothetical drug that prevents graft-versus-host disease.[19, 20] Use of the g-formula for our example was motivated by hypothesized time-varying confounding shown in the causal diagram in Figure 1, and by the desire to estimate the reduction in mortality from an intervention that could prevent graft-versus-host disease.

Study population: The study population arose from a multicenter trial of leukemia patients and comprises 137 individuals prepared for bone marrow transplants under a radiation-free regimen at four medical centers. Allogeneic bone marrow transplants were performed between March 1, 1984 and June 30, 1989, and patients were followed until death or administrative censoring at 5 years following transplant. Baseline covariates at time of transplant included age, sex, leukemia type (acute lymphocytic or acute myeloid leukemia), wait time from leukemia diagnosis to transplantation, and cytomegalovirus immune status (yes or no).

Patients were followed to assess when, if at all, platelets returned to the normal range (as a measure of immune function) and the patient experienced leukemia relapse.

We illustrate how to apply the parametric g-formula to the cohort of bone marrow transplant recipients to estimate the effect of a hypothetical intervention that prevents graft-versus-host disease from occurring. While the cohort is small, our example can be easily adapted to larger observational studies with long-term follow-up.

The parametric g-formula algorithm for the bone marrow transplant data: We use a 3-step algorithm for the parametric g-formula to estimate hazard ratios comparing: a) a new drug that prevents graft-versus-host disease from occurring during follow up ("prevented"); and b) no intervention ("natural course"). We compare "natural course" to "prevented" as a measure of how effective a drug that prevents graft-versus-host disease would have been at preventing mortality in our cohort (or similar populations). We present the algorithm in compact form in figure 2, and interested readers may find it of useful to follow our algorithm with the SAS code and data given in eAppendices 1-4 and the more formal (and technical) presentation of the g-formula of our data in eAppendix 5.

Step 1) *Probability modeling*: We start with a person-period data set in which each record corresponds to one person-day and time-varying covariates are represented as (0,1) dichotomous variables for each person day. Within this person-period data, we fit a pooled logistic model (i.e. a logistic model fit to all person periods) to estimate the log-odds of each of the time-varying covariates (graft-versus-host disease, platelet level, relapse, death), for each person period (i.e. the conditional log-odds of the covariate taking on the value "1").[21] We included time (i.e. days since transplant) in the model using a set of polynomial terms. Because we modeled the *onset* of graft-versus-host disease (and other time-varying covariates), the models for each covariate on day k were fit only using person-days for which the patient had not yet experienced each time-varying covariate on day $k-1$. We transformed the log-odds of each covariate into probabilities using the transformation $\Pr(\cdot) = \exp(\cdot)/(1+\exp(\cdot))$.

Step 2) *Monte Carlo sampling*: From our original data of N=137 patients, we re-sampled with replacement M=137,000 "pseudo-patients", retaining only baseline covariates. The sample should be as large as practical to minimize simulation error. This simulation is carried out as follows.

We simulated time-varying covariate and outcome data for each of the pseudo-patients using conditional probabilities generated in Step 1 and baseline covariates (with time-varying covariates set to 0 at baseline). Each pseudo-patient received values for these covariates on each day based on a draw from a Bernoulli distribution with the conditional mean given by the probability modeled in step 1, above. Essentially, on each day *k*, we flipped biased coins to choose whether or not the pseudo-patient became exposed, experienced relapse, returned to normal platelet levels, was censored, or died, and the probability of a coin coming up "heads" depended on values of those covariates on prior days and the baseline covariates.

The dataset from this step is known as the "natural course" because there is no intervention. To ensure that covariate distributions from the "natural course" closely follow distributions in the observed data, we examined the Kaplan-Meier survival curve for death (figure 4) as well as the complement of the Kaplan-Meier curves for time to onset of graft-versus-host disease, relapse, and normal platelet count (figure 3). We fit multiple parametric forms for each model in step 2 and based our final choice of model on how closely these graphs and covariate means from the "natural course" matched those in the observed data. For example, we compared fits of models using linear, quadratic, cubic polynomial, quadratic spline, unrestricted cubic spline, and restricted cubic splines. Model components and coefficient values are given in eAppendix 6.

We repeated Step 2 using an intervention: "prevented:" set graft-versus-host disease to 0 and generate all other covariates, forcing graft-versus-host disease to remain 0. By setting the values of graft-versus-host disease in the dataset, the proportion of deaths in the simulated dataset in which we set graft-versus-host disease=0 approximates the solution to the explicit formulas shown in our hypothetical example (in which the conditional probabilities of graft-versus-host disease onset are all set to 0).

Step 3) *Effect estimation*: We concatenated the datasets from step 2 estimated the hazard ratio comparing the hazards in the "natural course" dataset with those in the "prevented" dataset. This was done by using an indicator variable for the dataset (1="natural course", 0="prevented") and using that indicator as the exposure variable in a Cox model.

To estimate confidence intervals for the hazard ratio, we repeated Steps 1-3 on 4000 different samples of size 137 taken at random with replacement from the original data. The standard deviation (SD) of the 4000 log-hazard ratios approximates the standard error of the log-hazard ratio, and was used to calculate 95% confidence intervals (CI) using the normal approximation: log-hazard ratio ± 1.96*SD(log-hazard ratio).

Statistical methods for comparison

To compare the g-formula with standard methods, we estimated crude and covariate conditional hazard ratios (and 95% confidence intervals) for the effect of graft-versus-host disease on mortality using a Cox proportional hazards model for time-varying data (with observed data).[22] We controlled for possible confounding by baseline and time-varying covariates by including indicator terms in the Cox model. A test of proportional hazards in crude and regression-adjusted models indicated that a summary hazard ratio over the five-year course of the study was adequate.

RESULTS

Patients had a median age of 28 years (interquartile range: 21, 35), and 60% were male (Table 2). Half of the patients tested positive for cytomegalovirus at baseline (n=68; 50%). In the first 5 years of the study, 72 patients developed graft-versus-host disease and 43 patients relapsed. Platelet levels returned to normal for 88% of patients (n=120). Five years after receiving a transplant, over 58% of the patients had died (n=80).

Regression adjustment

The crude hazard ratio comparing the hazard of all-cause mortality among patients with and without graft-versus-host disease was 1.2 (95% CI: 0.77, 2.0; table 3). The hazard ratio was unchanged after regression adjustment for baseline covariates, but was notably larger after regression adjustment for baseline and time-varying covariates (hazard ratio: 2.3; 95% CI 1.4, 3.9).

G-formula

In the simulated data under the natural course, the cumulative distribution functions for platelet count, relapse, death, and graft-versus-host disease closely followed those in the observed data (Figures 3 and 4). The 5-year cumulative risk of death using the g-formula under the natural course was 61%. The hazard ratio comparing the “natural course” to “prevented” (referent) was 1.1 (95% CI: 0.91, 1.3).

DISCUSSION

In our analysis of survival data using a simple and yet easily adaptable application of the g-formula, we estimated the effects on mortality of an intervention to prevent graft-versus-host disease immediately after bone marrow transplant. Almost 60% of the bone marrow transplant patients in this cohort experienced graft-versus-host disease under the natural course; however, compared to a hypothetical intervention to prevent graft-versus-host disease, under the natural course we observed an increase in the relative mortality hazard of only 10%. Preventing graft-versus-host disease does not appear to markedly reduce mortality risk because other factors that influence the risk of subsequent mortality, such as leukemia relapse, are not decreased by the hypothetical intervention to prevent graft-versus-host disease. Rather, preventing graft-versus-host disease may increase the rate of relapse: 42% of “pseudo-individuals” experienced relapse under the “prevented” intervention, whereas 33% experienced relapse under the “natural course.” These observations agree with the typically coincident occurrences of graft-versus-host disease and graft-versus-leukemia, in which donor cells attack residual leukemia cells in the transplant recipient and may help to prevent relapse. Graft-versus-host disease has been observed to correlate with a lower rate of leukemia relapse and is hypothesized to reduce the probability of relapse through immune mediation processes, of which normal platelet count is an indicator.[1]

Cox regression analysis with adjustment for time-varying covariates suggested a substantially higher mortality hazard for those with graft-versus-host disease when compared to those without graft-versus-host disease. Because the Cox model cannot appropriately control time-varying confounding when the confounders are causal intermediates, we hypothesize that the Cox model overestimates the total effect of graft-versus host disease on mortality. Therefore, the difference between the g-formula estimate and the regression adjustment method may be due, in part, to over-adjustment, which introduces bias into our estimate of the total effect of graft-versus-host disease.[23] In addition, another reason for the difference between the g-formula estimate and the adjusted Cox regression model estimate is that these analyses estimate different quantities. The g-formula yielded an estimate of the hazard ratio comparing the observed mortality in our cohort (the natural course) with the expected mortality in that cohort under the new treatment. The Cox model yielded an estimate of the mortality hazard for those with graft-versus-host disease compared to those without graft-versus-host disease. [24]Using the g-formula to estimate of the mortality hazard if everyone had graft-versus-host disease compared to the expected mortality under the new treatment resulted in a hazard ratio intermediate between our original estimate and the adjusted Cox model estimate (hazard ratio=1.8; 95%CI: 0.94, 3.3). This contrast is a closer analogue to the adjusted Cox model hazard ratio than our original g-formula estimate, but it does not address the study question that motivated the analysis.

The g-formula can be used to estimate risk ratios or differences, which are easier to interpret and less subject to some biases than are hazard ratios.[25] Like Westreich et al. (2012), we have estimated hazard ratios as effect measures to ease the comparison with results from conventional methods.[14] However, when a confounder is a strong predictor of the outcome, marginal hazard ratios from the g-formula and conditional hazard ratios from regression approaches may differ even if there is no time-varying confounding, a condition known as "non-collapsibility."[26] Because the set of time-varying covariates includes relapse, a strong predictor of mortality, this is not a trivial concern for our example, but adjustment via regression modeling simply replaces confounding bias with bias due to conditioning on an effect of the exposure. However, the g-formula estimates a hazard ratio we could observe in a population intervention (marginal, or standardized to the population), and is arguably more useful than a conditional hazard ratio when estimating the public health impact of interventions.

There is little knowledge about how potential time-varying confounding by relapse and platelet levels affects estimates of excess mortality due to graft-versus-host disease. G-methods (the body of methods that derive from the g-formula, including structural nested models and marginal structural models) may provide an avenue towards a clearer understanding. G-methods have previously been shown to yield results congruent with clinical trial data (where time-varying confounding may be minimized) in situations where conventional analyses of observational data yield incongruous estimates.[14, 27, 28] There are few examples in the literature in which the parametric g-formula has been used to estimate the effects of policies or interventions. Robins (1986, 1987) estimated the effect of interventions capping arsenic exposure on the risk of lung cancer in a cohort 8,047 copper smelter workers in Montana.[8, 10] Ahern et al. (2009) used the parametric g-formula to estimate the effect of interventions to change neighborhood smoking norms on the prevalence of smoking using data collected from 4000 New York City residents.[11] Using data from the Nurses' Health Study, Taubman et al. (2009) estimated the 20-year risk (cumulative incidence) of coronary heart disease under interventions on dietary factors, exercise, smoking, alcohol consumption, and body mass index;[12] using data from the same study, Danaie et al. (2013) estimated the effects of similar interventions on the 24-year risk of type-2 diabetes[16] and Garcia-Aymerich et al. estimated the effect of joint interventions on physical activity and weight loss on adult onset asthma incidence.[17] In a cohort of 8392 HIV-infected participants in the HIV-CAUSAL collaboration, Young et al. (2011) applied the parametric g-formula to estimate 5-year mortality risks under seven dynamic treatment regimes to determine the optimal CD4 count at which to begin antiretroviral therapy.[13] Westreich et al. (2012) used the parametric g-formula to estimate hazard ratios comparing the hazard of AIDS or death among 1498 HIV-infected patients enrolled in the Multicenter AIDS Cohort Study and the Women's Interagency HIV Study had all study participants received antiretroviral therapy with the hazard had none of the study participants received therapy.[14] Finally, using data from a cohort of 3,002 textile workers, Cole et al. (2013) estimated effect on lung cancer mortality if recent occupational limits on annual exposures to chrysotile asbestos fibers had been in place during the workers' tenure.[15]

In all cases except the study by Ahern et al., the potential for time-varying confounding precluded the use of conventional regression approaches. In all cases, the g-formula naturally lent itself to estimating the effects of potential interventions. Rather than estimating the more familiar contrast in measures of disease occurrence for a unit change in the exposure, the g-formula provides an estimate of outcome occurrence under a specific treatment regime under study. While regression adjustment (and our g-formula comparison of mortality under always/never graft-versus-host disease interventions) allow for estimation of "etiologic" hazard ratios, only the g-formula easily allows estimation of the effect of preventing graft-versus-host disease in the population, which may be more useful for informing population level interventions.

We must make several assumptions when using of the g-formula to estimate the effects of exposure on an outcome in observational data.[8] We provide a brief discussion of these assumptions for our analysis, which are reviewed in depth elsewhere.[29, 30] These assumptions are not unique to the g-formula, but estimating effect measures in the g-formula requires explicitly confronting these assumptions, which may be valuable in informing the interpretation of observational studies.

Conditional exchangeability: As epidemiologists, we may strive to measure and control for strong confounders of the exposure-outcome relationship to avoid making inference based on spurious relationships, and we often assume we have been successful – this is the assumption of conditional exchangeability. In our g-formula example, we impute the potential outcomes for each individual based on an evolving covariate history generated by predictive models for the exposure, covariates and the outcome. If there were a strong, unmeasured, baseline confounder, we would have omitted an important variable from both the exposure model and the outcome model. However, as Robins et al note, the exposure model does not need to be correct in order to make inference from the g-formula.[18] To see this, recall our original hypothetical example in which, to make interventions, the probability of exposure is set to 0 in the "prevent" intervention. Thus, because we "set" exposure, the hazard under some intervention will be subject to bias only through the model for the outcome. The practical consequence of this is that unmeasured confounder bias should be (approximately) the same under the g-formula and a standard regression model, when we are comparing two interventions. However, unmeasured confounding bias may be greater in the g-formula when comparing an intervention versus the "natural course," since this analysis requires a model for the exposure.

In more complex settings, such as if the omitted confounder were also a cause of another time-varying covariate, the g-formula may be subject to more bias than conventional models due to unmeasured confounding. Intuitively, this is because we could be accumulating bias over multiple models, rather than a single outcome model (as in the standard Cox model). However, we should note that, in complex time-varying settings, conventional models require stratification over time-varying confounders, which may be problematic for reasons discussed above, and conventional models may not be able to estimate useful effect measures, such as the impact of an intervention on a population. The g-formula is not subject to either of these shortcomings. Ultimately, the impact of unmeasured confounding in the g-formula (versus conventional models) is an open question; future work will provide guidance to epidemiologists working in settings where unmeasured confounding is expected to be problematic.

While sensitivity analyses exist for examining robustness to the assumption of no unmeasured confounding,[31] they are most informative when one can hypothesize both a source of confounding and the plausible levels of the strength of the confounder associations in the study.[32] We know of no factors that could strongly influence the estimated effect of Graft-versus-host disease on mortality, so in the case of our g-formula example, the sensitivity of the hazard ratio to unmeasured confounding would either be purely speculative or would be focused on finding the level of confounding that could explain the observed association.

Treatment version irrelevance: We assume the effect of exposure is the same whether we set it, as in the case of step 5 of our algorithm or a clinical trial, or if it occurs naturally.[33, 34] When we "intervene" to set graft-versus-host disease to 0 for all pseudo-patients in our Monte Carlo sample, treatment version irrelevance means we assume that this emulates a process in which a researcher could prevent graft-versus-host disease. We hypothesize this assumption may be violated in our example, because acute and chronic forms of graft-versus-host disease may not affect mortality with the same magnitude.[2] In this case, we could improve on how well we meet this assumption by including models for both chronic and acute graft-versus-host disease, as well as estimating effects for both on mortality. However, because we are estimating the population averaged-effect of graft-versus-host disease on mortality, such an analysis would not likely give substantially different results. Our analysis applies to graft-versus-host disease, as it occurs in populations similar to our cohort in which graft-versus-host disease can be either acute or chronic. Consistency requires that we specify interventions which are unambiguous.[35] Thus, the sensitivity of g-formula results to violations of treatment version irrelevance depends, in some respect, on how well one can emulate one's analysis with a clinical trial.

As an anonymous referee pointed out, our analysis can be used to guide future clinical trials by informing on the effects of a drug that completely prevents graft-versus-host disease – our estimate of a 10% reduction in 5-year mortality could used to plan such a study. If one were to study the effect of say, a population intervention that dictated an instantaneous shift in BMI category, then g-formula results would not estimate the effect we could observe in any clinical trial.[35]

Correct model specification: The parametric g-formula is especially vulnerable to the assumption of correct model specification, due to the use of multiple models. As an example from our study, if the true effect of wait time to transplant had a quadratic association with exposure and a cubic association with mortality, then we would have misspecified two models and would likely have introduced bias into the hazard ratio. Our g-formula algorithm allows informal checking of this assumption by comparing the observed data to the data simulated under the natural course (e.g. figures 3 and 4). Nonetheless, similarity to the natural course cannot completely rule out model misspecification.[18] The g-formula may not be ideal for hypothesis generating studies, in which causal relationships are more uncertain and model misspecification is potentially severe.

Summary, recommendations and conclusions

The g-formula does not require more information than standard methods to estimate effects of exposures or interventions and is not subject to bias due to stratifying on variables affected by exposure. The g-formula does not assume that the hazard ratio is homogenous over the levels of the confounders – that is, we are not restricted to estimating a summary hazard ratio that conditions on the covariates. Unlike methods that use stratification of the data by covariates, and then obtain a summary hazard ratio under the assumption of a constant hazard ratio over levels of the confounder, the g-formula permits us to obtain a summary hazard ratio without the without stratifying the effect measure by over covariates. This relaxed assumption comes at the cost of wider confidence intervals, so there is a bias-variance trade-off to consider when deciding if the g-formula is an appropriate statistical tool for an epidemiologic analysis. If one knows that time-varying confounding is not present, or that exposure could not affect subsequent confounders, then standard regression methods may be more appropriate due mainly to concerns about model misspecification in the g-formula. However, if time-varying confounding is possible, use of the g-formula provides some assurance that one is not introducing bias by inappropriately stratifying on effects of exposure, and it is useful to check results from standard regression models against those from the g-formula, as we have shown.

We could have estimated the hazard ratio for the marginal effect of graft-versus-host disease on mortality using inverse-probability weighted marginal structural models or adaptations of structural nested models.[36-38] In principle, these models would yield similar effect estimates as our method, though our small data set would likely result in practical violations of additional assumptions necessary for inverse probability weighting methods,[39] and, without knowledge about the baseline hazard of potential outcomes, structural nested models may yield larger confidence intervals.[40] These methods may be more desirable when model misspecification is a primary concern, because they require fewer models than the g-formula.[18]

Though the g-formula was first described in 1986 by Robins, the availability of rich data from large cohort studies and the acceleration of computing speeds have only recently made the method feasible for widespread use. While the g-formula may be more sensitive to model misspecification than standard regression models, we show how inappropriate control of time-varying confounding could lead to incorrect conclusions about the strength of the effect of Graft-versus-host disease on mortality. Moreover, we illustrate how the g-formula allows estimation of the effects of realistic interventions and give concrete examples of situations in which the g-formula should be used by epidemiologists. We have provided both publicly available data and the SAS code as appendices to this paper. The current analysis represents the first published use of the g-formula for a time-varying exposure that can be easily replicated by other investigators using a real-world example. Future implementations of the g-formula can extend our example easily to an array of epidemiologic problems and further explore the impacts on inference from unmeasured confounding and model misspecification.

References:

1. Sullivan K, Weiden P, Storb R, Witherspoon R, Fefer A, Fisher L, Buckner C, Anasetti C, Appelbaum F, and Badger C. Influence of acute and chronic graft-versus-host disease on relapse and survival after bone marrow transplantation from HLA-identical siblings as treatment of acute and chronic leukemia [published erratum appears in Blood 1989 Aug 15; 74 (3): 1180]. *Blood*. 1989; 73:1720-1728.

2. Horowitz MM, Gale RP, Sondel PM, Goldman JM, Kersey J, Kolb HJ, Rimm AA, Ringdén O, Rozman C, and Speck B. Graft-versus-leukemia reactions after bone marrow transplantation. *Blood*. 1990; 75:555-562.

3. Robins JM and Wasserman L. Estimation of effects of sequential treatments by reparameterizing directed acyclic graphs. *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*. 1997; 409-420.

4. Keiding N, Filiberti M, Esbjerg S, Robins JM, and Jacobsen N. The graft versus leukemia effect after bone marrow transplantation: A case study using structural nested failure time models. *Biometrics*. 1999; 55:23-28.

5. Hernán MA and Robins JM. Causal Inference. Boca Raton, FL: Chapman & Hall/CRC; 2013.

6. Rosenbaum PR. The consquences of adjustment for a concomitant variable that has been affected by the treatment. *J R Stat Soc Ser A-G*. 1984; 147:656-666.

7. Weinberg CR. Toward a clearer definition of confounding. *Am J Epidemiol*. 1993; 137:1-8.

8. 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; 7:1393-1512.

9. Hernán MA and Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Commun H*. 2006; 60:578-586.

10. Robins JM. A graphical approach to the identification and estimation of causal parameters in mortality studies with sustained exposure periods. *J Chron Dis*. 1987; 40:139S-161S.

11. Ahern J, Hubbard A, and 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; 169:1140-1147.

12. Taubman SL, Robins JM, Mittleman MA, and Hernán MA. Intervening on risk factors for coronary heart disease: an application of the parametric g-formula. *Int J Epidemiol*. 2009; 38:1599-1611.

13. Young JG, Cain LE, Robins JM, O'Reilly EJ, and Hernán MA. Comparative effectiveness of dynamic treatment regimes: an application of the parametric g-formula. *Stat Biosci*. 2011; 3:119-143.

14. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, Gange SJ, and Hernán MA. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat Med*. 2012; 31:2000-2009.

15. Cole SR, Richardson DB, Chu H, and Naimi AI. Analysis of Occupational Asbestos Exposure and Lung Cancer Mortality Using the G Formula. *Am J Epidemiol*. 2013; 177:989-996.

16. Danaei G, Pan A, Hu FB, and Hernán MA. Hypothetical midlife interventions in women and risk of type 2 diabetes. *Epidemiology*. 2013; 24:122-128.

17. Garcia-Aymerich J, Varraso R, Danaei G, Camargo CA Jr, and 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; 179:20-6.

18. Robins JM, Hernán MA, and Siebert U. Effects of multiple interventions. In: Comparative Quantification of Health Risks: The Global and Regional Burden of Disease Attributable to Major Risk Factors. Geneva: World Health Organization; 2004.

19. Copelan EA, Biggs JC, Thompson JM, Crilley P, Szer J, Klein JP, Kapoor N, Avalos BR, Cunningham I, and Atkinson K. Treatment for acute myelocytic leukemia with allogeneic bone marrow transplantation following preparation with BuCy2. *Blood*. 1991; 78:838-843.

20. Klein JP and Moeschberger ML. Examples of Survival Data. In: *Survival Analysis: Techniques for Censored and Truncated Data*. 2nd ed. New York: Springer; 2003:1-21.

21. D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, and Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham Heart Study. *Stat Med*. 1990; 9:1501-15.

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

23. Schisterman EF, Cole SR, and Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009; 20:488-95.

24. Cole SR and Hernán MA. Fallibility in estimating direct effects. *Int J Epidemiol*. 2002; 31:163-5.

25. Hernán MA. The hazards of hazard ratios. *Epidemiology*. 2010; 21:13-15.

26. Greenland S. Absence of confounding does not correspond to collapsibility of the rate ratio or rate difference. *Epidemiology*. 1996; 7:498-501.

27. Cole SR, Hernán MA, Robins JM, Anastos K, Chmiel JS, Detels R, Ervin C, Feldman J, Greenblatt RM, Kingsley L, and others. Effect of highly active antiretroviral therapy on time to acquired immunodeficiency syndrome or death using marginal structural models. *Am J Epidemiol*. 2003; 158:687-694.

28. Hernán MA, Cole SR, Margolick J, Cohen M, and Robins JM. Structural accelerated failure time models for survival analysis in studies with time-varying treatments. *Pharmacoepidem Dr S*. 2005; 14:477-491.

29. Cole SR and Hernán MA. Constructing inverse probability weights for marginal structural models. *Am J Epidemiol*. 2008; 168:656-64.

30. Daniel RM, Cousens SN, De Stavola BL, Kenward MG, and Sterne JAC. Methods for dealing with time-dependent confounding. *Stat Med*. 2013; 32:1584-618.

31. Robins JM. Marginal structural models versus structural nested models as tools for causal inference. Statistical Models in Epidemiology: The Environment and Clinical Trials. *The Environment and Clinical Trials*. *Halloran MS, Berry D, eds*. *IMA*. 1999; 116:95--134.

32. Lash TL, Fox MP, and Fink AK. Applying quantitative bias analysis to epidemiologic data. New York: Springer Verlag; 2009

33. Cole SR and Frangakis CE. The consistency statement in causal inference: a definition or an assumption?. *Epidemiology*. 2009; 20:3-5.

34. Hernán MA and VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology*. 2011; 22:368-377.

35. Hernán MA and Taubman SL. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *Int J Obes (Lond)*. 2008; 32 Suppl 3:S8-14.

36. Robins JM. The analysis of randomized and non-randomized AIDS treatment trials using a new approach to causal inference in longitudinal studies. In: Sechrest L, Freeman H, Mulley A, eds. Health service research methodology: a focus on AIDS. *Health service research methodology: a focus on AIDS*. 1989; 113-159:113-159.

37. Robins JM and Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time-dependent covariates. *Biometrika*. 1992; 79:311-319.

38. Robins JM, Hernán MA, and Brumback BA. Marginal structural models and causal inference in epidemiology. *Epidemiology*. 2000; 11:550-560.

39. Westreich D and Cole SR. Invited commentary: positivity in practice. *Am J Epidemiol*. 2010; 171:674-677.

40. Joffe MM, Yang WP, and Feldman H. G-estimation and artificial censoring: problems, challenges, and applications. *Biometrics*. 2012; 68:275-286.

Figure Legends

Figure 1: Directed acyclic graph showing hypothesized causal relationships between study variables for days $k-1$ and k and demonstrating bias in regression stratification methods in estimating the effect of exposure over time ($GvHD_{k-1}, GvHD_k$) on subsequent death when time-varying factors on confounding pathways ($Relapse_k$) may be affected by prior exposure.

Figure 2: Parametric g-formula algorithm for bone marrow transplant data.

Figure 3: Incidence curves for death (lower left) return to normal platelet levels (top left) relapse (top right) and graft-versus-host-disease (GvHD, bottom right) from both observed (gray line) and g-formula natural course Monte Carlo (black line) data.

Figure 4: Survival functions a) observed from the bone marrow transplant data b) from the natural course intervention in the g-formula c) from two hypothetical interventions "never Graft-versus-host disease" (top line) "always Graft-versus-host disease" (bottom line) after bone marrow transplants using the g-formula. The gray line indicates the observed survival curve, while the solid black lines indicate the survival curves from the Monte Carlo data for the g-formula interventions (from top) never Graft-versus-host disease, Natural Course, always Graft-versus-host disease.

Table1

Table 1: Estimating risk under interventions using the g-formula with the hypothetical example given in methods section of the main text.

G-formula component		Value
Conditional probabilities	**First 5 years**	PR(exposed): 0.5 Risk among exposed: 0.025 Risk among unexposed: 0.025
	Last 5 years	Pr(exposed): 0.5 Risk among exposed: 0.060 Risk among unexposed: 0.051
Risk under interventions	**Observed/ Natural course**	0.051*(1/2)*(1-0.025*(1/2)) + 0.025*(1/2) + 0.060*(1/2)*(1-0.025*(1/2)) + 0.025*(1/2) = 0.080
	Always exposed	0.051*(0)*(1-0.025*(0)) + 0.025*(0) + 0.060*(1)*(1-0.025*(1)) + 0.025*(1) = 0.084
	Never exposed	0.051*(1)*(1-0.025*(1)) + 0.025*(1) + 0.060*(0)*(1-0.025*(0)) + 0.025*(0) = 0.075
	Expose if survive to year 5	0.051*(0)*(1-0.025*(0)) + 0.025*(1/2) + 0.060*(1)*(1-0.025*(1)) + 0.025*(1/2) = 0.084

Table2

Table 2. Characteristics of 137 patients receiving bone marrow transplants during treatment for leukemia at 4 study sights between 1985 and 1989.

		No.	**%**
Sex	Male	80	58
	Female	57	42
GvHD onset within 5 years	No	65	47
	Yes	72	53
Relapse within 5 years	No	94	69
	Yes	43	31
Return to normal platelet count within 5 years	No	17	12
	Yes	120	88
Leukemia type	ALL	38	28
	Low risk AML	54	39
	High risk AML	45	33
CMV status at baseline	Negative	69	50
	Positive	68	50
Vital status 5 years after transplant	Alive or censored	57	42
	Dead	80	58
		Median	**IQR**
Continuous variables	Age (years)	28	21, 35
	Days to transplant	6	4, 8
	Follow up	547	183, 1377
	Days to relapse	467	122, 1363
	Days to normal platelets	18	14, 27

Table3

Table 3. Hazard ratios (HR) comparing the hazard of all-cause mortality between patients with and without graft-versus-host disease [a] (regression) or comparing cohorts with and without hypothetical intervention to prevent graft-versus-host disease (g-formula)

Method	HR	95% CI
Regression		
Crude	1.2	0.77, 2.0
Baseline adjusted [b]	1.2	0.71, 1.9
Fully adjusted [c]	2.3	1.4, 3.9
G-formula		
Natural course vs. Prevent [d]	1.1	0.91, 1.3

[a] Graft vs. Host Disease

[b] Baseline covariates include age at date of bone marrow transplant, wait time until transplant, sex, and cytomegalovirus status at baseline.

[c] Adjusted for baseline covariates above and time-varying covariates, including days during which platelets had not returned to normal, cumulative days the patient had not experienced relapse, and indicators for relapse and platelets returning to normal on a given day.

[d] Comparing the hazard of all-cause mortality between the entire cohort simulated under no intervention and the entire cohort of simulated to be unexposed (referent) at all time points.

Fig1

$$\begin{tikzcd}
GvHD_{k-1} \arrow[r] \arrow[rr, bend left] & Relapse_k \arrow[r] \arrow[d] & Death_k \\
 & GvHD_k \arrow[ur] &
\end{tikzcd}$$

Fig2

Step 1: Model conditional probabilities in observed data

Using a person-period data set (108714 person-days):

A) Model: $Pr(platelet_k = 1 | GvHD_{k-1}, platelet_{k-1}, relapse_{k-1}, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$

B) Model: $Pr(relapse_k = 1 | GvHD_{k-1}, platelet_k, relapse_{k-1}, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$

D) Model: $Pr(censored_k = 1 | GvHD_k, platelet_k, relapse_k, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$

E) Model: $Pr(death_k = 1 | GvHD_k, platelet_k, relapse_k, baseline\ variables, death_{k-1} = censored_k = 0)$

Step 2: Generate time-varying exposures, covariates, and outcomes in Monte Carlo sample

Sample with replacement N=137,000 “pseudo-patients” from baseline data only to simulate follow-up from baseline variables and conditional probabilities and set $GvHD_0 = 0, platelet_0 = 0, relapse_0 = 0$

For “pseudo-patient” $i = 1, 2..., 137000; Day\ k = 1, 2, ...,$ day of death or administrative censoring

Using conditional probabilities from step 1:

F) Generate $platelet_k$ using $Pr(platelet_k = 1 | GvHD_{k-1}, platelet_{k-1}, relapse_{k-1}, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$

G) Generate $relapse_k$ using $Pr(relapse_k = 1 | GvHD_{k-1}, platelet_k, relapse_{k-1}, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$

H) Generate $GvHD_k$ using $Pr(GvHD_k = 1 | GvHD_{k-1}, platelet_k, relapse_k, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$
If intervening, set $GvHD_k = 1$ (Always exposed) or $GvHD_k = 0$ (Never exposed)

I) Generate $Pr(censored_k = 1 | GvHD_k, platelet_k, relapse_k, baseline\ variables, death_{k-1} = censored_{k-1} = 0)$
If intervening, set $censored_k = 0$ (Never lost to follow-up)

J) Using $Pr(death_k = 1 | GvHD_k, platelet_k, relapse_k, baseline\ variables, death_{k-1} = censored_k = 0)$:

Generate $death_k$	If $death_k = 0$ and not administrative censoring time, go to day k+1, repeat F-J
	If $death\ _k$= 1 or admin. censoring time, go to “pseudo-patient” $i + 1$. If end of data set, go to step 3

Step 3: Estimate effect measure

Concatenate data sets from 2 interventions from step 2 (e.g. natural course, always exposed, never exposed)

Fit statistical model to Monte Carlo data to compare outcome proportions/rates between 2 data sets

Fig3

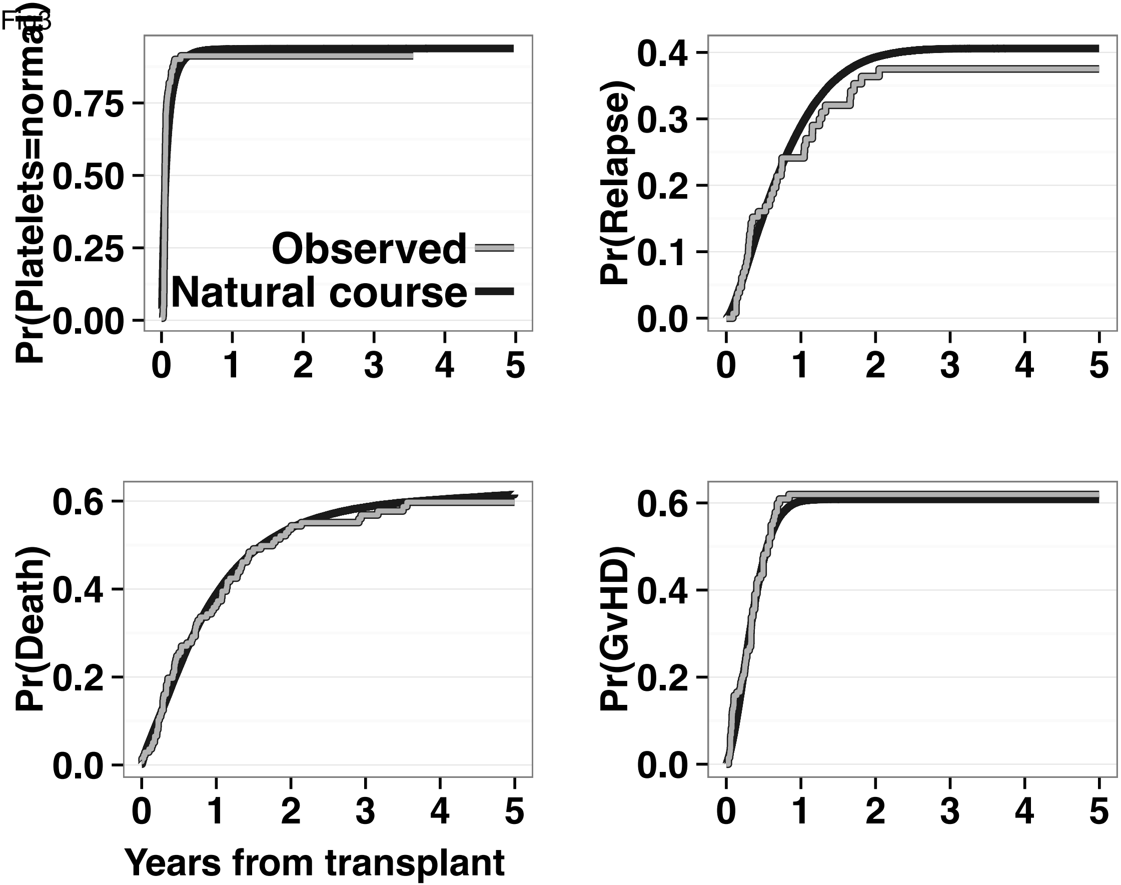
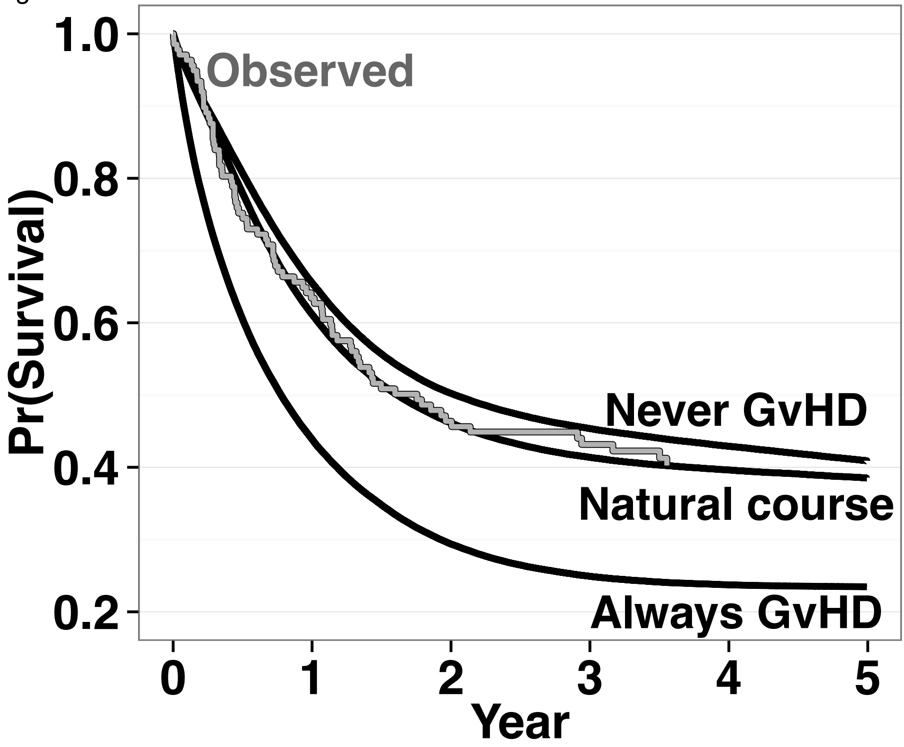

Pr(Platelets=normal)
0.00
0.25
0.50
0.75
Observed
Natural course
0
1
2
3
4
5
Pr(Relapse)
0.0
0.1
0.2
0.3
0.4
0
1
2
3
4
5
Pr(Death)
0.0
0.2
0.4
0.6
0
1
2
3
4
5
Years from transplant
Pr(GvHD)
0.0
0.2
0.4
0.6
0
1
2
3
4
5

Fig4


Observed
Never GvHD
Natural course
Always GvHD
Pr(Survival)
1.0
0.8
0.6
0.4
0.2
0
1
2
3
4
5
Year

Appendices

Appendix 1: SAS code for generating person-period data from bone marrow transplant data;

```
*) Step 1 - generate person-day data from bone marrow transplant data;
DATA person_day_level;
 SET person_level;
 BY id;

 *initial values for time-varying variables;
 daysnorelapse=0;daysnoplatnorm=0;daysnogvhd=0;
 daysrelapse=0;daysplatnorm=0;daysgvhd=0;

 *time-varying variables;
 DO day = 1 TO t;
  yesterday = day-1;
  daysq = day**2;
  daycu = day**3;
  *cubic spline, day, (knots=83.6 401.4 947.0 1862.2);
  daycurs1 = ((day>83.6)*((day-83.6)/83.6)**3)+((day>1862.2)*((day-1862.2)/83.6)**3)*(947.0-83.6) -((day>947.0)*((day-947.0)/83.6)**3)*(1862.2-83.6)/(1862.2-947.0);
  daycurs2 = ((day>401.4)*((day-401.4)/83.6)**3)+((day>1862.2)*((day-1862.2)/83.6)**3)*(947.0-401.4) -((day>947.0)*((day-947.0)/83.6)**3)*(1862.2-401.4)/(1862.2-947.0);
  d =   (day>=t)*d_dea;
  gvhd =   (day>t_gvhd);
  relapse =  (day>t_rel);
  platnorm =  (day>t_pla);
  *lagged variables;
  gvhdm1 = (yesterday>t_gvhd);
  relapsem1 =  (yesterday>t_rel);
  platnormm1 =  (yesterday>t_pla);
 censeof = 0; censlost=0;
  IF day = t AND d = 0 THEN DO;
   IF day = 1825 THEN censeof = 1;
   ELSE censlost=1;
  END;

  IF relapse = 0 THEN daysnorelapse + 1;
  IF platnorm = 0 THEN daysnoplatnorm + 1;
  IF gvhd = 0 THEN daysnogvhd + 1;
  IF relapse = 1 THEN daysrelapse + 1;
  IF platnorm = 1 THEN daysplatnorm + 1;
  IF gvhd = 1 THEN daysgvhd + 1;

  KEEP id age: male cmv day: yesterday d relapse: platnorm: gvhd: all censlost wait;
  OUTPUT;
 END;
RUN;
```

Appendix 2: SAS code for generating model coefficients for use in G-formula (model coefficient values given in appendix 6)

```
*Step 2) - estimate modeling coefficients used to generate probabilities;
TITLE "Parametric G-formula coefficient estimation models";
PROC LOGISTIC DATA = person_day_level DESC;
 TITLE2 "Model for probability of relapse=1 at day k";
 WHERE relapsem1=0;
 MODEL relapse = all cmv male age gvhdm1 daysgvhd platnormm1 daysnoplatnorm agecurs1
agecurs2 day
 daysq wait;
 ODS OUTPUT ParameterEstimates=rmod(KEEP=variable estimate);*keep model coefficients;
PROC LOGISTIC DATA = person_day_level DESC;
 TITLE2 "Model for probability of platnorm=1 at day k";
 WHERE platnormm1=0;
 MODEL platnorm = all cmv male age agecurs1 agecurs2 gvhdm1 daysgvhd daysnorelapse wait;
 ODS OUTPUT ParameterEstimates=Pmod(KEEP=variable estimate);*keep model coefficients;
PROC LOGISTIC DATA = person_day_level DESC;
 TITLE2 "Model for probability of exposure=1 at day k";
 WHERE gvhdm1=0;
 MODEL gvhd = all cmv male age platnormm1 daysnoplatnorm relapsem1 daysnorelapse
agecurs1 agecurs2
 day daysq wait;
 ODS OUTPUT ParameterEstimates=gmod(KEEP=variable estimate);*keep model coefficients;
PROC LOGISTIC DATA = person_day_level DESC;
 TITLE2 "Model for probability of censoring=1 at day k";
 MODEL censlost = all cmv male age daysgvhd daysnoplatnorm daysnorelapse agesq day
daycurs1 daycurs2
 wait;
 ODS OUTPUT ParameterEstimates=cmod(KEEP=variable estimate); *keep model coefficients;
PROC LOGISTIC DATA = person_day_level DESC;
 TITLE2 "Model for probability of outcome=1 at day k";
 MODEL d = all cmv male age gvhd platnorm daysnoplatnorm relapse daysnorelapse agesq day
daycurs1
 daycurs2 wait day*gvhd daycurs1*gvhd daycurs2*gvhd ;
 ODS OUTPUT ParameterEstimates=dmod(KEEP=variable estimate);*keep model coefficients;
RUN;

*create data sets with coefficients with prefixes p(platnorm) r(relapse) g(gvhd) c(censoring)
d(death);
DATA Pmod(DROP=i j variable estimate);
 SET Pmod END=eof;
 j+1;
 ARRAY p[11];
 RETAIN p:;
 DO i= 1 TO j; IF i = j THEN p[i] = estimate; END;
 IF eof THEN OUTPUT;
DATA Rmod(DROP=i j variable estimate);
 SET Rmod END=eof;
 j+1;
 ARRAY r[14];
 RETAIN r:;
 DO i= 1 TO j; IF i = j THEN r[i] = estimate; END;
 IF eof THEN OUTPUT;
DATA Gmod(DROP=i j variable estimate);
 SET Gmod END=eof;
 j+1;
 ARRAY g[14];
 RETAIN g:;
 DO i= 1 TO j; IF i = j THEN g[i] = estimate; END;
 IF eof THEN OUTPUT;
DATA Cmod(DROP=i j variable estimate);
 SET Cmod END=eof;
 j+1;
 ARRAY c[13];
 RETAIN c:;
 DO i= 1 TO j; IF i = j THEN c[i] = estimate; END;
 IF eof THEN OUTPUT;
DATA Dmod(DROP=i j variable estimate);
 SET Dmod END=eof;
 j+1;
 ARRAY d[18];
 RETAIN d:;
 DO i= 1 TO j;IF i = j THEN d[i] = estimate; END;
 IF eof THEN OUTPUT;
RUN;
*merge model coefficient values into PERSON LEVEL data;
DATA person_level_w_coefs;
 SET person_level;
 IF _N_=1 THEN DO;
  SET pmod;
  SET gmod;
  SET rmod;
  SET dmod;
  SET cmod;
 END;
RUN;
```

Appendix 3: Drawing Monte Carlo sample and running natural course / GvHD intervention using G-formula

```
*Step 3) - sample with replacement from data;
PROC SURVEYSELECT DATA=person_level_w_coefs SEED=12131231 OUT=mcsample
METHOD=URS N=137000 OUTHITS;
RUN;

*Step 4 and 5) - run Monte Carlo sample for natural course, always and never GvHD;
DATA natcourse(KEEP = id all cmv male age d td gvhd tg platnorm tp relapse tr)
  alwaysgvhd(KEEP = id all cmv male age d td gvhd tg platnorm tp relapse tr)
  nevergvhd(KEEP = id all cmv male age d td gvhd tg platnorm tp relapse tr);
   SET mcsample; *set each time the intervention changes;
   BY id;
   CALL STREAMINIT(187100);
 DO intervention = 0 TO 2;
  * RETAIN done 0;
   day = 0;
   done = 0;
 DO WHILE (day <= 1825 AND done=0);
   day+1;
   daysq = day**2;
   daycu = day**3;
   *cubic spline, day, (knots=83.6 401.4 947.0 1862.2);
   daycurs1 = ((day>83.6)*((day-83.6)/83.6)**3)+((day>1862.2)*((day-
1862.2)/83.6)**3)*(947.0-83.6) -((day>947.0)*((day-947.0)/83.6)**3)*(1862.2-83.6)/(1862.2-
947.0);
   daycurs2 = ((day>401.4)*((day-401.4)/83.6)**3)+((day>1862.2)*((day-
1862.2)/83.6)**3)*(947.0-401.4) -((day>947.0)*((day-947.0)/83.6)**3)*(1862.2-
401.4)/(1862.2-947.0);
   IF day =1 THEN DO; *set baseline variables;
    relapse=0;gvhd=0;platnorm=0;
    gvhdm1=0;relapsem1=0;platnormm1=0;
    daysnorelapse=0;daysnoplatnorm=0;daysnogvhd=0;
    daysrelapse=0;daysplatnorm=0;daysgvhd=0;
   END;*set baseline variables;
   ELSE DO;*set time-varying variables - lag is built in;
    IF relapse = 0 THEN daysnorelapse + 1;
 ELSE daysrelapse + 1;
    IF platnorm = 0 THEN daysnoplatnorm + 1;
    ELSE daysplatnorm + 1;
    IF gvhd = 0 THEN daysnogvhd + 1;
    ELSE daysgvhd + 1;
    *platelets (Time-varying covariate L1);
    IF platnormm1=1 THEN platnorm=1; *assume platelets stay normal once they reach normal
levels;
    ELSE DO; *normal platelet probability at day k;
    logitpp = p1 + p2*all + p3*cmv + p4*male + p5*age + p6*agecurs1 + p7*agecurs2 +
p8*gvhdm1 + p9*daysgvhd + p10*daysnorelapse + p11*wait;
    IF logitpp <-700 THEN gvhd = 1;*avoid machine error;
    ELSE platnorm=RAND("bernoulli",1/(1+exp(-(logitpp))));
   END; *normal platelet probability at day k;
   *relapse(Time-varying covariate L2);
   IF relapsem1=1 THEN relapse=1; *assume relapse is not cured once patient experiences first
post transplant relapse;
   ELSE DO; *relapse probability at day k;
    logitpr= r1 + r2*all + r3*cmv + r4*male + r5*age + r6*gvhdm1 + r7*daysgvhd +
r8*platnormm1 + r9*daysnoplatnorm + r10*agecurs1 + r11*agecurs2 + r12*day + r13*daysq +
r14*wait;
    IF logitpr <-700 THEN relapse = 1;  *avoid machine error;
    ELSE relapse=RAND("bernoulli",1/(1+exp(-(logitpr))));
   END;*relapse probability at day k;
   END;
   *GvHD (main exposure A);
   IF gvhdm1=1 THEN gvhd=1; *assume patients can't be cured of GvHD once GvHD onset
occurs;
   ELSE DO; *gvhd onset probability at day k;
    logitpg = g1 + g2*all + g3*cmv + g4*male + g5*age + g6*platnormm1 + g7*daysnoplatnorm
+ g8*relapsem1 + g9*daysnorelapse + g10*agecurs1 + g11*agecurs2 + g12*day + g13*daysq +
g14*wait;
    IF logitpG <-700 THEN gvhd = 1;  *avoid machine error;
    ELSE gvhd = RAND("bernoulli",1/(1+exp(-(logitpg))));
   END;*gvhd onset probability at day k;

   *intervene on exposure;
   IF intervention = 0 THEN gvhd=gvhd; *natural course;
   ELSE IF intervention = 1 THEN gvhd=1; *always GvHD;
   ELSE IF intervention = 2 THEN gvhd=0; *never GvHD;

   IF done=0 THEN DO; *censoring and death probability at day k;
    *censoring probability at day k;
    logitpc =  c1 + c2*all + c3*cmv + c4*male + c5*age + c6*daysgvhd + c7*daysnoplatnorm +
c8*daysnorelapse + c9*agesq + c10*day + c11*daycurs1 + c12*daycurs2 + c13*wait;
    IF logitpc <-700 THEN d = 1;  *avoid machine error;
    ELSE cens = RAND("bernoulli",1/(1+exp(-(logitpc))));
    IF intervention > 0 THEN cens=0; *intervening to prevent censoring for everything but
natural course;
    done=cens;
    IF done=0 THEN DO; *if not censored on day k;
     *death probability at day k;
     logitpd = d1 + d2*all + d3*cmv + d4*male + d5*age + d6*gvhd + d7*platnorm +
d8*daysnoplatnorm + d9*relapse + d10*daysnorelapse + d11*agesq + d12*day + d13*daycurs1
+ d14*daycurs2 + d15*wait + d16*day*gvhd + d17*daycurs1*gvhd + d18*daycurs2*gvhd;
```

```
      IF logitpd <-700 THEN d = 1;*avoid machine error;
      ELSE d = RAND("bernoulli",1/(1+exp(-(logitpd))));
      done=d;
     END;*if not censored on day k;
     IF day >= 1825 THEN done=1;
     IF gvhd=1 AND gvhdm1=0 THEN tg=day;
     IF relapse=1 AND relapsem1=0 THEN tr = day;
     IF platnorm=1 AND platnormm1=0 THEN tp = day;
     IF done=1 THEN DO;
      td=day;
      IF gvhd=0 THEN tg=day+1;
      IF relapse=0 THEN tr=day+1;
      IF platnorm=0 THEN tp=day+1;
      IF intervention = 0 THEN OUTPUT natcourse; *output a PERSON LEVEL dataset;
      ELSE IF intervention = 1 THEN OUTPUT alwaysgvhd; *output a PERSON LEVEL dataset
if intervention is always GvHD;
      ELSE IF intervention = 2 THEN OUTPUT nevergvhd; *output a PERSON LEVEL dataset if
intervention is never GvHD;
    END;*censoring and death probability at day k;
   END;*set time-varying variables;
   *lagged variables;
   relapsem1=relapse;
   platnormm1=platnorm;
   gvhdm1=gvhd;
  END; * while done = 0 and day < 1825;
 END;*intervetion from 0 to 2;
RUN;

*Step 6) concatentate intervetion data sets and run Cox model;
DATA gformula;
 SET alwaysgvhd nevergvhd;

PROC PHREG DATA = gformula;
 MODEL td*d(0) = gvhd / TIES=EFRON RL;
RUN;

PROC PHREG DATA = gformula;
 MODEL td*d(0) = gvhd1 gvhd2 / TIES=EFRON RL;
 gvhd1=gvhd*(td<=100); gvhd2=gvhd*(td>100);
RUN;
```

Appendix 4: SAS code to read bone marrow transplant data;

```
DATA person_level;
 INPUT   id t t_rel d_dea t_gvhd d_gvhd d_rel t_pla d_pla age male cmv waitdays all ;
```

```
DATALINES;
1 1 1 1 1 0 0 1 0 42 1 0 196 1
2 2 2 1 2 0 0 2 0 20 1 0 75 0
3 10 10 1 10 0 0 10 0 34 1 1 240 0
4 16 16 1 16 0 0 16 0 27 0 1 180 0
5 35 35 1 35 0 0 35 0 23 0 1 150 0
6 48 48 1 48 0 0 14 1 32 0 1 150 0
7 53 53 1 53 0 0 53 0 33 0 1 180 0
8 62 47 1 62 0 1 11 1 27 1 0 90 0
9 63 63 1 38 1 0 16 1 44 1 0 360 0
10 73 64 1 73 0 1 38 1 45 0 1 180 0
11 74 74 1 29 1 0 24 1 41 0 1 750 0
12 79 79 1 16 1 0 79 0 43 0 0 90 0
13 80 80 1 10 1 0 80 0 30 0 0 150 0
14 80 80 1 21 1 0 0 1 35 1 0 780 0
15 86 86 1 86 0 0 86 0 17 1 1 239 1
16 93 47 1 93 0 1 28 1 7 1 0 135 0
17 97 76 1 97 0 1 97 0 48 1 1 330 0
18 105 105 1 21 1 0 15 1 37 1 1 120 0
19 105 105 1 105 0 0 105 0 14 1 0 150 0
20 105 48 1 105 0 1 30 1 17 0 0 210 0
21 107 107 1 107 0 0 107 0 30 1 1 178 1
22 110 74 1 110 0 1 49 1 28 1 1 303 1
23 121 100 1 28 1 1 65 1 39 1 1 210 0
24 122 122 1 88 1 0 13 1 20 1 0 2616 1
25 122 120 1 122 0 1 12 1 25 0 1 510 0
26 128 115 1 128 0 1 12 1 37 0 1 270 0
27 129 93 1 129 0 1 51 1 37 0 1 240 0
28 153 113 1 153 0 1 59 1 31 0 1 240 0
29 156 104 1 28 1 1 20 1 20 1 0 85 1
30 162 109 1 162 0 1 40 1 36 1 1 393 1
31 162 162 1 162 0 0 13 1 22 1 0 300 0
32 164 164 1 164 0 0 164 0 19 0 0 285 0
33 168 168 1 168 1 0 48 1 32 0 1 150 0
34 172 172 1 22 1 0 37 1 40 0 0 129 1
35 183 183 1 130 1 0 21 1 11 0 0 120 0
36 194 194 1 94 1 0 25 1 26 0 0 329 1
37 195 32 1 195 0 1 16 1 36 1 0 90 0
38 222 219 1 123 1 1 52 1 28 1 1 120 0
39 226 226 0 226 0 0 10 1 18 0 0 208 1
40 243 122 1 243 0 1 23 1 37 0 1 170 1
41 248 157 1 100 1 1 52 1 33 0 1 180 0
42 262 192 1 10 1 1 59 1 29 1 1 74 1
43 262 55 1 262 0 1 24 1 23 0 1 331 1
44 265 242 1 210 1 1 14 1 32 1 0 180 0
45 269 110 1 120 1 1 27 1 29 0 1 361 1
```

```
46 276 276 1 81 1 0 21 1 18 0 0 146 1
47 288 288 1 18 1 0 288 0 45 1 1 90 0
48 318 318 1 140 1 0 12 1 35 0 1 300 0
49 341 268 1 21 1 1 17 1 20 0 1 180 0
50 350 332 1 350 0 0 33 1 22 1 0 834 1
51 363 363 1 363 0 0 19 1 52 1 1 180 0
52 371 230 1 184 1 1 9 1 39 0 0 147 1
53 390 390 1 390 0 0 11 1 50 1 0 120 0
54 392 273 1 122 1 1 24 1 43 1 1 240 0
55 393 381 1 100 1 1 16 1 33 0 0 120 0
56 414 414 1 414 0 0 27 1 21 1 0 120 0
57 417 383 1 417 0 1 16 1 15 1 0 824 1
58 418 418 1 220 1 0 21 1 18 1 0 110 1
59 431 272 1 431 0 1 12 1 30 0 1 120 0
60 466 466 1 119 1 0 100 1 15 1 0 508 1
61 469 467 1 90 1 1 20 1 35 0 1 120 0
62 481 481 1 30 1 0 24 1 35 1 1 90 0
63 487 487 1 76 1 0 22 1 22 1 0 128 1
64 491 422 1 180 1 1 491 0 22 0 0 210 0
65 515 390 1 515 0 1 31 1 23 1 1 210 0
66 522 421 1 25 1 1 20 1 28 1 0 90 0
67 526 526 1 121 1 0 11 1 15 1 0 943 1
68 530 530 0 38 1 0 34 1 17 1 0 151 1
69 547 456 1 130 1 1 24 1 31 1 1 630 0
70 583 486 1 583 0 1 11 1 17 0 0 120 0
71 641 641 1 641 0 0 11 1 26 1 0 90 0
72 653 211 1 653 0 1 23 1 23 1 0 90 0
73 677 677 1 150 1 0 8 1 15 1 1 150 0
74 704 704 1 36 1 0 18 1 29 0 1 105 0
75 716 662 1 716 0 1 17 1 28 1 0 84 1
76 732 625 1 732 0 1 18 1 39 0 1 150 0
77 781 609 1 781 0 1 26 1 27 1 1 187 1
78 845 845 0 845 0 0 20 1 40 0 1 210 0
79 847 847 0 847 0 0 16 1 28 1 0 75 0
80 848 848 0 155 1 0 16 1 23 1 0 180 0
81 860 860 0 860 0 0 15 1 25 0 0 180 0
82 932 932 0 29 1 0 7 1 27 0 0 60 0
83 957 957 0 957 0 0 69 1 18 1 0 90 0
84 996 996 0 72 1 0 12 1 22 1 0 1319 1
85 1030 1030 0 210 1 0 14 1 25 0 0 210 0
86 1063 1063 1 240 1 0 16 1 50 1 1 270 0
87 1074 1074 1 120 1 0 19 1 30 1 1 150 0
88 1111 1111 0 1111 0 0 22 1 19 1 0 236 1
89 1136 1136 0 140 1 0 15 1 47 1 1 900 0
90 1156 748 1 180 1 1 18 1 14 1 0 60 0
91 1167 1167 0 39 1 0 1167 0 27 0 1 191 1
```

```
92 1182 1182 0 112 1 0 22 1 24 0 0 203 1
93 1199 1199 0 91 1 0 29 1 24 1 0 174 1
94 1238 1238 0 250 1 0 18 1 24 1 1 240 0
95 1258 1258 0 120 1 0 66 1 30 0 1 180 0
96 1279 129 1 1279 0 1 22 1 17 0 0 937 1
97 1298 84 1 1298 0 1 1298 0 8 0 1 105 0
98 1324 1324 0 25 1 0 15 1 46 1 1 75 0
99 1330 1330 0 96 1 0 17 1 20 1 1 1006 1
100 1345 1345 0 32 1 0 14 1 50 1 1 120 0
101 1356 606 0 1356 0 1 14 1 33 1 1 210 0
102 1363 1363 0 200 1 0 12 1 13 1 1 90 0
103 1377 1377 0 123 1 0 12 1 22 1 1 2187 1
104 1384 1384 0 200 1 0 19 1 21 0 0 120 0
105 1433 1433 0 236 1 0 12 1 32 1 1 93 1
106 1447 1447 0 220 1 0 24 1 33 0 1 150 0
107 1462 1462 0 70 1 0 13 1 17 0 0 168 1
108 1470 1470 0 180 1 0 14 1 27 1 0 240 0
109 1496 1496 0 307 1 0 12 1 26 1 1 127 1
110 1499 248 0 1499 0 1 9 1 35 1 0 30 0
111 1527 1527 0 1527 0 0 13 1 22 0 0 450 0
112 1535 1535 0 1535 0 0 21 1 35 0 0 180 0
113 1562 1562 0 1562 0 0 18 1 26 1 1 90 0
114 1568 1568 0 1568 0 0 14 1 15 1 0 90 0
115 1602 1602 0 139 1 0 18 1 21 1 0 1720 1
116 1631 1631 0 150 1 0 40 1 27 1 1 690 0
117 1674 1674 0 1674 0 0 24 1 37 1 0 60 0
118 1709 1709 0 20 1 0 19 1 23 0 1 90 0
119 1799 1799 0 140 1 0 12 1 32 1 0 120 0
120 1825 1825 0 1825 0 0 19 1 19 1 1 210 0
121 1825 1825 0 1825 0 0 19 1 34 0 1 270 0
122 1825 1825 0 1825 0 0 9 1 37 0 0 180 0
123 1825 1825 0 260 1 0 15 1 29 0 1 90 0
124 1825 1825 0 230 1 0 16 1 33 0 1 225 0
125 1825 1825 0 180 1 0 16 1 35 0 0 105 0
126 1825 1825 0 67 1 0 13 1 26 1 1 98 1
127 1825 1825 0 250 1 0 17 1 36 0 0 240 0
128 1825 1825 0 220 1 0 18 1 27 1 1 210 0
129 1825 1825 0 1825 0 0 12 1 25 0 0 60 0
130 1825 1825 0 1825 0 0 11 1 16 1 1 60 0
131 1825 1825 0 52 1 0 15 1 45 0 0 105 0
132 1825 1825 0 150 1 0 17 1 35 1 0 120 0
133 1825 1825 0 1825 0 0 16 1 35 1 1 120 0
134 1825 1825 0 1825 0 0 14 1 29 1 0 24 0
135 1825 1825 0 1825 0 0 17 1 31 1 0 60 0
136 1825 1825 0 1825 0 0 21 1 19 1 1 270 0
137 1825 1825 0 1825 0 0 22 1 18 1 0 750 0
```

```
;

*define more baseline covariates;
DATA person_level;
 SET person_level;
 *baseline variables;
 wait = waitdays/30.5;
 agesq = age**2;
 *restricted cubic spline on age (knots at 17, 25.4, 30, 41.4);
 agecurs1 = (age>17.0)*(age-17.0)**3-((age>30.0)*(age-30.0)**3)*(41.4-17.0)/(41.4-30.0);
 agecurs2 = (age>25.4)*(age-25.4)**3-((age>41.4)*(age-41.4)**3)*(41.4-25.4)/(41.4-30.0);
RUN;
```

Appendix 5:

Formal treatment of the parametric G-formula

We adopt the "do notation" of Pearl (2009) to express the potential outcomes and covariate values we would observe under interventions, where, for example $X\big(do(B = b)\big)$ is the value we would expect *X* to take on if we could intervene on *B* by setting it to the value "*b*" (i.e. uppercase letters are random variables, and lowercase letters are realizations). While all covariates are recorded at the individual level, we omit the subscript *i* for clarity. Bold font is used to denote vectors. Time is subscripted and history variables are denoted with an overbar, where a history variable $\bar{B}_k = (B_0, \dots, B_k)$, is an expanding set of time-specific random variables or a summary of the history (such as the cumulative time on treatment) through time *k*. Notation for our data is shown in Appendix Table 1.

The G-formula for the bone marrow transplant data

Using our data, the G-formula for the marginal incidence of death by the end of follow up at time *j* (*I_j*) under no intervention can be expressed as:

$$I_j$$

$$= \sum_{k=1}^{j} \sum_{v} \sum_{l} \sum_{a} \left\{ \begin{array}{l} Pr(Y_k = 1 |\, A_k = a_k, \bar{A}_k = \bar{a}_k, \boldsymbol{L_k} = \boldsymbol{l_k}, \bar{\boldsymbol{L}}_{\boldsymbol{k}} = \bar{\boldsymbol{l}}_{\boldsymbol{k}}, \boldsymbol{V_0} = \boldsymbol{v_0}, Y_{k-1} = 0) \times \\ \prod_{m=1}^{k} \left[\begin{array}{l} Pr\big(C_m = 0 \big|\, A_m = a_m, \bar{A}_m = \bar{a}_m, \boldsymbol{L_m} = \boldsymbol{l_m}, \bar{\boldsymbol{L}}_{\boldsymbol{m}} = \bar{\boldsymbol{l}}_{\boldsymbol{m}}, \boldsymbol{V_0} = \boldsymbol{v_0}, Y_{m-1} = C_{m-1} = 0\big) \times \\ Pr\big(A_m = a_m \big|\, \bar{A}_{m-1} = \bar{a}_{m-1}, \boldsymbol{L_m} = \boldsymbol{l_m}, \bar{\boldsymbol{L}}_{\boldsymbol{m}} = \bar{\boldsymbol{l}}_{\boldsymbol{m}}, \boldsymbol{V_0} = \boldsymbol{v_0}, Y_{m-1} = C_{m-1} = 0\big) \times \\ Pr\big(L_m = l_m \big|\, A_{m-1} = a_{m-1}, \bar{A}_{m-1} = \bar{a}_{m-1}, \bar{\boldsymbol{L}}_{\boldsymbol{m}} = \bar{\boldsymbol{l}}_{\boldsymbol{m}}, \boldsymbol{V_0} = \boldsymbol{v_0}, Y_{m-1} = C_{m-1} = 0\big) \times \\ Pr(V_0 = v_0) \times \\ Pr(Y_{m-1} = 0 |\, A_{m-1} = a_{m-1}, \bar{A}_{m-1} = \bar{a}_{m-1}, \boldsymbol{L_{m-1}} = \boldsymbol{l_{m-1}}, \bar{\boldsymbol{L}}_{\boldsymbol{m-1}} = \bar{\boldsymbol{l}}_{\boldsymbol{m-1}}, \boldsymbol{V_0} = \boldsymbol{v_0}, Y_{m-2} = C_{m-2} = 0) \end{array} \right] \end{array} \right\}$$

Where time specific conditional probability of Y_k at each time point is multiplied by the conditional probabilities (or probability densities) of $A_{k,}$, $\boldsymbol{L_k}$, and $\boldsymbol{V_0}$ and summed over the observed values of $a_{k,}$, $\boldsymbol{l_k}$ and $\boldsymbol{v_0}$. Because $\boldsymbol{V_0}$ includes only baseline covariates, it is not assumed to be a function of any other variables of interest. With enough data and if all variables were low dimension (e.g. dichotomous), we could identify the time specific probability of each variable in the set $(Y_k, C_k, A_k, \bar{A}_{k-1}, \boldsymbol{L_k}, \bar{\boldsymbol{L}}_{\boldsymbol{k-1}}, V_0)$ simply by taking sample proportions (i.e. the model is non-parametrically identified). However, this cohort is followed for 1825 days and we include continuous covariates, so we must model each probability.

3. Parametric G-formula algorithm

Step 1) From a one-record-per-person data set recording baseline covariates, and time to: relapse; normal platelet count; GvHD; and death or censoring, create a person-period data set in which each record corresponds to one person-day (for example, the third person in the data set given appendix 4 who survives to day 10 has 10 records in the dataset). Each person-period record contains a variable *k* recording the number of days since transplant as well as the set of variables $(Y_k, C_k, A_k, \bar{A}_k, \boldsymbol{L_k}, \bar{\boldsymbol{L}}_{\boldsymbol{k}}, V_0)$.

Step 2) Use a pooled logistic model to estimate the conditional probability of each of the time-varying covariates, including the exposure. When the time-specific conditional probability of the covariate is low (say $\Pr(Y_k = 1| \cdot) < 0.1$), coefficients from a pooled logistic model with flexible terms for time (e.g., spline) closely approximate those of a Cox model (Abbot 1985). We used the SAS procedure LOGISTIC to output model coefficients. For our example data, the pooled logistic model to estimate the probability of developing GvHD at time *k* (the second line in the G-formula above) was

$$\Pr(A_k = 1 |\, \bar{A}_{k-1} = 0, \boldsymbol{L_k}, \boldsymbol{\bar{L}_k}, \boldsymbol{V_0}, Y_{k-1} = C_{k-1} = 0; \alpha) =$$

$$expit(\sum \hat{\alpha}_G \boldsymbol{G_k} + \sum \hat{\alpha}_L \boldsymbol{L_k} + \sum \hat{\alpha}_{\bar{L}} \boldsymbol{\bar{L}_k} + \sum \hat{\alpha}_V \boldsymbol{V_0})$$

and the logistic model for death at time *k* was

$$\Pr(Y_k = 1 |\, A_k, \bar{A}_{k-1}, \boldsymbol{L_k}, \boldsymbol{\bar{L}_k}, \boldsymbol{V_0}, Y_{k-1} = C_{k-1} = 0; \omega) =$$

$$\mathrm{expit}(\sum \widehat{\omega}_G \boldsymbol{G_k} + \widehat{\omega}_A A_k + \widehat{\omega}_{\bar{A}} \bar{A}_{k-1} + \sum \widehat{\omega}_{GA} \boldsymbol{G_k} A_k + \sum \widehat{\omega}_L \boldsymbol{L_k} + \sum \widehat{\omega}_{\bar{L}} \boldsymbol{\bar{L}_k} + \sum \widehat{\omega}_V \boldsymbol{V_0})$$

Where $\mathrm{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$ is the inverse-logit function, $\boldsymbol{G_k}$ is a vector of terms representing a restricted cubic spline for days on study (day *k*) with knots at the 10th, 40th, 60th and 90th percentiles (83, 401, 947 and 1862 days), and the $\hat{\alpha}$ (or $\widehat{\omega}$) coefficients represent the difference in log-odds of GvHD onset (or death) on day *k* for a one unit difference in each covariate. The hat accent over the coefficients is to emphasize that we are estimating these coefficients from the data. The baseline covariates in $\boldsymbol{V_0}$, the current and prior time-varying confounder vectors $\boldsymbol{L_k}$ and $\boldsymbol{\bar{L}_k}$, and current and prior GvHD variables A_k and $\bar{A}_k$ are as described in the notation section, and the summation symbol Σ is used to indicate that each covariate within a vector will be associated with a unique coefficient. Additionally $\sum \boldsymbol{G_k} A_k$ is set of product terms between GvHD and the spline variables for time to allow for changing direct exposure effects over time. Using the terminology of directed acyclic graphs, the set of covariates $(\bar{A}_k, \boldsymbol{L_k}, \boldsymbol{\bar{L}_k}, \boldsymbol{V_0})$ should be a sufficient set of covariates such that conditioning on them blocks all non-causal pathways from exposure to the outcome (Pearl 2009 Ch 4).

Because we are modeling the onset of GvHD, the model for A_k is conditioned on $\bar{A}_{k-1} = 0$ (the patient is GvHD free up until just before day *k*) by restricting the logistic model to the subset of the data where $\bar{A}_{k-1} = 0$. The person-period data contain observations only for the days from the first day after transplant to the day of death or censoring, so all of our models are implicitly conditioned on $Y_{k-1} = C_{k-1} = 0$.

We also used pooled logistic models for $\boldsymbol{L_k} = (L_{1k}, L_{2k})$ where $L_{1k}$is platelet level and L_{2k} is relapse, where

$$\Pr(L_{1k} = 1|\, A_{k-1}, \bar{A}_{k-1}, \bar{L}_{1k-1} = 0, \bar{L}_{2k-1}, \boldsymbol{V_0}, Y_{k-1} = C_{k-1} = 0; \beta) =$$

$$expit\left(\sum\hat{\beta}_G \boldsymbol{G_k} + \hat{\beta}_A A_{k-1} + \hat{\beta}_{\bar{A}} \bar{A}_{k-1} + \hat{\beta}_{\bar{L}2} \bar{L}_{2k-1} + \sum\hat{\beta}_V \boldsymbol{V_0}\right)$$

and

$$\Pr(L_{2k} = 1|A_{k-1}, \bar{A}_{k-1},\ L_{1k}, \bar{L}_{1k-1}, \bar{L}_{2k-1} = 0, \boldsymbol{V_0}, Y_{k-1} = C_{k-1} = 0; \gamma) =$$

$$expit(\sum\hat{\gamma}_G \boldsymbol{G_k} + \hat{\gamma}_A A_{k-1} + \hat{\gamma}_{\bar{A}} \bar{A}_{k-1} + \hat{\gamma}_{L1} L_k + \hat{\gamma}_{\bar{L}1} \bar{L}_{1k-1} + \sum\hat{\gamma}_V \boldsymbol{V_0})$$

As in the logistic model for GvHD, $\boldsymbol{G_k}$ is a flexible function of time, $\boldsymbol{V_0}$ are the baseline covariates, and we model the return of normal platelet counts or relapse by conditioning on $\bar{L}_{1k-1} = 0$ (or $\bar{L}_{2k-1} = 0$). $\bar{L}_{1k-1}$ and $\bar{L}_{2k-1}$ are the days spent without normal platelet counts or without relapsing. The $\hat{\beta}$ (or $\hat{\gamma}$) parameters represent the difference in the log odds of return to normal platelet counts (or relapse) on day *k* for a one-unit increment of the corresponding covariate. We assume that, in a given day the temporal order is $(L_{1k}, L_{2k}, A_k, C_k, Y_k)$. The log-odds of censoring was assumed to be a linear function of baseline covariates, cumulative days with abnormal platelet counts, cumulative days spent relapse-free, and cumulative days with GvHD.

Step 3) From our original sample of N=137, we re-sampled with replacement M=137,000 pseudo-patients, retaining only baseline covariates $\boldsymbol{V_0}$. The large sample reduces Monte Carlo error, and should be as large as is practical. Resampling can be done, for example, using the SAS procedure SURVEYSELECT.

Step 4) Using model coefficients generated in Step 2 and the baseline covariates from our 137,000 pseudo-patients, we generated follow-up data for each of the M pseudo-patients by imputing values for platelet levels, relapse, and graph-versus host disease. Time-varying covariates at baseline were set to A_0=0, and $\boldsymbol{L_0}$=(0,0). We also imputed the outcome variable Y_1, using observed baseline covariates and the imputed values for A_1 and $\boldsymbol{L_1}$. Similar to the dataset created in step one, we retained a record for each of the 137,000 pseudo-patients for each person-day. For example, the value of A_k (the indicator of GvHD on day k) for individuals who were previously GvHD-free and not yet censored or dead was generated from a binomial distribution with

$$\Pr(A_k = 1 | \boldsymbol{L_k}, \boldsymbol{\bar{L}_{k-1}}, \boldsymbol{V_0}) =$$

$$expit(\sum \hat{\alpha}_G \boldsymbol{G_k} + \sum \hat{\alpha}_L \boldsymbol{L_k} + \sum \hat{\alpha}_{\bar{L}} \boldsymbol{\bar{L}_{k-1}} + \sum \hat{\alpha}_V \boldsymbol{V_0})$$

Step 2 can be performed in SAS with a single DATA step using DO loops to cycle through days 1 to 1825 (or until $Y_k = 1$ or $C_k = 1$), and the GvHD values can be imputed for each person day by drawing a value from a Bernoulli distribution with the probability of GvHD onset ($\Pr(A_k = 1 | \boldsymbol{L_k}, \boldsymbol{\bar{L}_{k-1}}, \boldsymbol{V_0})$) given above. As was observed in our example data, we set this probability to 1 if the pseudo-patient developed GvHD on a previous day.

Exposure, covariate, censoring and outcome values for each subsequent day were imputed in the same way, using imputed covariate values from previous days (e.g. day *k*-1) to generate new values for subsequent days. Any pseudo-patient with $Y_k = 1$ or $C_k = 1$ did not receive subsequent records for times *k*+1, …, 1825.

Rather than model the distribution of the baseline covariates in $\boldsymbol{V_0}$ from which we could have generated baseline covariates values, we used the joint empirical distribution of the baseline covariates. To do this, we kept the baseline covariate values from our original data (N) and used them to generate time-varying covariate values for days $k > 1$ in our pseudo data (M). With this Monte Carlo dataset we checked marginal survival curves and covariate distributions against those from the observed data. Model selection was carried out by repeating Steps 1 and 2, and varying the parametric forms (e.g., …) of each model until the marginal survival curves and covariate distributions in the Monte Carlo data (*M*) closely approximated those in the observed data (*N*). We refer to the data *M* generated from this set of models as the "natural course."

Step 5) We repeated Step 4 using two interventions: a) "Always GvHD:" set GvHD to 1 on day 1 and impute all other covariates as before, and b) "Never GvHD:" set GvHD to 0 and impute all other covariates, not allowing GvHD status to change. In both interventions, we set $C_k = 0$ for all censoring other than the end of follow up. With no drop out and no competing risks, we could estimate *E[Y(do(A$_k$ =1))]* and *E[Y(do(A$_k$ =0))]* by simply taking sample proportions of the deaths in each simulated dataset.

For example, the model to impute the return to normal platelet counts for the data for the intervention $do(A_k = 1)$ is expressed as:

$$\Pr(L_{1k} = 1 | \, A_{k-1}, \bar{A}_{k-2}, \bar{L}_{2k-1}, \boldsymbol{V_0}; \beta) =$$

$$expit\left(\sum\hat{\beta}_G\boldsymbol{G_k} + \hat{\beta}_A A_{k-1} + \hat{\beta}_{\bar{A}}\bar{A}_{k-1} + \hat{\beta}_{\bar{L}2}\bar{L}_{2k-1} + \sum\hat{\beta}_V\boldsymbol{V_0}\right) =$$

$$expit\left(\sum\hat{\beta}_G\boldsymbol{G_k} + \hat{\beta}_A 1 + \hat{\beta}_{\bar{A}}(k-1) + \hat{\beta}_{\bar{L}2}\bar{L}_{2k-1} + \sum\hat{\beta}_V\boldsymbol{V_0}\right)$$

Where GvHD (A_{k-1}) is always 1 and days since onset of GvHD ($\bar{A}_{k-1}$) is the number of days since transplant. The intervention is carried out across all four models (i.e. the models for return to normal platelet count, relapse, GvHD, and death) and yields data, the distribution of which corresponds to what we would observe in the population of bone marrow transplant had we been able to implement the intervention $do(A_k = 1)$ (i.e. give all patients GvHD immediately after surgery). We repeat this process setting GvHD to 0 for every day k.

Step 6) We concatenated the datasets from step 5 and fit a marginal structural Cox proportional hazards model to the simulated dataset to estimate the HR comparing the hazard of *Y(do(A_k =1))* to the hazard of *Y(do(A_k =0))*, which, under assumptions outlined later can be interpreted as a causal HR. The marginal structural Cox model for the potential failure times $T\left(do(A_k = a)\right)$ can be expressed as:

$$\lambda_{k1}^{*} = \lambda_{k0}^{*}(\exp(\eta A_k))$$

and, for a = 1 or 0

$$\lambda_{ka}^{*} = \lim_{\delta k \to 0}\left(\frac{\Pr\left(k < T\left(do(A_k = a)\right) < k + \delta k \middle| T\left(do(A_k = a)\right) > k\right)}{\delta k}\right)$$

Where $T\left(do(A_k = a)\right)$ is the day on which death occurred and the hazards λ_{k1}^{*} and λ_{k0}^{*} for the potential outcomes we would observe under the interventions *do(A_k =1)* and *do(A_k =0)*. Because the generated data from step 5 correspond to the data we would see under these two interventions, a marginal Cox model (a model in which exposure is the only independent variable) estimates the contrast between the interventions. To allow for non-proportional hazards, we also fit a Cox model to estimate separate HRs for the periods 0-100 days and 101-1825 days. As an estimate of the impact of an intervention to prevent GvHD in our cohort, we also estimated the HR comparing the hazards $Y(A_k)$ and $Y(do(A_k =0))$, where $Y(A_k)$ is the set of outcomes in the natural course data.

Step 7) To estimate confidence intervals for the HR, we repeated Steps 1-6 on 2000 different samples of size 137 taken at random with replacement from the original data N. The standard deviation of the 2000 log HRs approximates the standard error of the log HR, and was used to calculate 95% Wald bootstrap confidence intervals.

Appendix 6: Notation and model coefficients from predictive models in step 2

Appendix Table 1. Variable notation for the study of 137 patients receiving bone marrow transplants during treatment for leukemia at 4 study sights between 1985 and 1989.

Variable	Elements
Y_k	indicator of death (1= yes, 0=no) at the end of day *k* after bone marrow transplant
A_k	indicator of GvHD (1= yes, 0=no) at the end of day *k* after bone marrow transplant
$\bar{A}_k$	number of days since onset of GvHD (or 0 if onset has not occurred) as of the end of day *k*
$\boldsymbol{L_k}$	vector of observed indicators of 1) relapse or 2) normal platelet levels (1=patient has relapsed or reached normal platelet count, 0=not in relapse or below normal platelets) at the end of day *k* after bone marrow transplant
$\boldsymbol{\bar{L}_k}$	vector of 1) observed history of relapse or 2) normal platelet levels (1= patient has relapsed or reached normal platelet count prior to day *k*, 0=not in relapse or below normal platelets) prior to day *k* and 3) time (in days) spent relapse-free or 4) time spent without reaching normal platelet levels (i.e. these variables count up from day one until relapse or normal platelet levels are reached, after which they remain fixed) up to the end of day *k* after bone marrow transplant
$\boldsymbol{V_0}$	age, sex, leukemia type (acute lymphocytic or acute myeloid leukemia), wait time from leukemia diagnosis to transplantation, and cytomegalovirus immune status (yes or no)
C_k	indicator of censoring due to loss-to-follow up at time *k*
$Y_k(do(A_k=a_k))$	indicator of *potential* death (1= yes, 0=no) at the end of day *k* after bone marrow transplant, had we been able to intervene on GvHD and set it to the value a_k (i.e. we could either give a patient GvHD or prevent it)

Appendix Table 1: Predictive pooled logistic model coefficients for relapse on day k in a cohort of 137 bone marrow transplant patients. Parameter names correspond to variable names given in SAS code from appendix 2.

Parameter	Estimate	Std. Error	Chi-Square	p
Intercept	-6.868	1.308	27.558	<0.001
all	0.587	0.391	2.248	0.134
cmv	0.559	0.335	2.793	0.095
male	-0.254	0.351	0.524	0.469
age	-0.090	0.051	3.158	0.076
gvhdm1	-0.303	0.487	0.387	0.534
daysgvhd	-0.001	0.002	0.397	0.529
platnormm1	1.173	0.799	2.158	0.142
daysnoplatnorm	0.005	0.002	5.002	0.025
agecurs1	0.000	0.000	5.303	0.021
agecurs2	0.000	0.000	3.929	0.048
day	0.002	0.002	0.931	0.335
daysq	0.000	0.000	3.263	0.071
wait	-0.009	0.017	0.255	0.614

Appendix Table 2: Predictive pooled logistic model coefficients for return to normal platelet count on day k in a cohort of 137 bone marrow transplant patients. Parameter names correspond to variable names given in SAS code from appendix 2.

Parameter	Estimate	Std. Error	Chi-Square	p
Intercept	-5.772	0.640	81.447	<0.001
all	-0.071	0.233	0.093	0.761
cmv	-0.599	0.197	9.255	0.002
male	0.361	0.206	3.073	0.080
age	0.106	0.027	15.057	0.000
agecurs1	0.000	0.000	3.393	0.066
agecurs2	0.000	0.000	0.085	0.771
gvhdm1	-1.111	0.813	1.865	0.172
daysgvhd	-0.014	0.017	0.692	0.406
daysnorelapse	-0.005	0.004	1.717	0.190
wait	0.013	0.008	2.569	0.109

Appendix Table 3: Predictive pooled logistic model coefficients for graph-versus-host disease onset on day k in a cohort of 137 bone marrow transplant patients. Parameter names correspond to variable names given in SAS code from appendix 2.

Parameter	Estimate	Std. Error	Chi-Square	p
Intercept	-7.251	1.003	52.210	<0.001
all	0.601	0.287	4.382	0.036
cmv	0.105	0.257	0.166	0.684
male	-0.147	0.271	0.294	0.588
age	-0.002	0.041	0.003	0.955
platnormm1	0.430	0.476	0.817	0.366
daysnoplatnorm	0.010	0.007	2.044	0.153
relapsem1	0.087	1.553	0.003	0.955
daysnorelapse	0.091	0.107	0.722	0.396
agecurs1	0.000	0.000	0.992	0.319
agecurs2	0.000	0.000	1.016	0.314
day	-0.080	0.107	0.553	0.457
daysq	0.000	0.000	7.606	0.006
wait	0.013	0.010	1.824	0.177

Appendix Table 4: Predictive pooled logistic model coefficients for non-administrative censoring on day k in a cohort of 137 bone marrow transplant patients. Parameter names correspond to variable names given in SAS code from appendix 2.

Parameter	Estimate	Std. Error	Chi-Square	p
Intercept	-9.887	2.242	19.446	<0.001
all	0.759	0.448	2.873	0.090
cmv	-0.434	0.348	1.556	0.212
male	0.223	0.370	0.364	0.546
age	-0.080	0.107	0.559	0.455
daysgvhd	0.001	0.000	2.436	0.119
daysnoplatnorm	0.001	0.001	0.343	0.558
daysnorelapse	0.000	0.001	0.069	0.793
agesq	0.001	0.002	0.398	0.528
day	-0.001	0.006	0.012	0.913
daysq	0.000	0.000	1.132	0.287
daycu	0.000	0.000	1.891	0.169
wait	-0.004	0.012	0.118	0.732

Appendix Table 5: Predictive pooled logistic model coefficients for mortality on day k in a cohort of 137 bone marrow transplant patients. Parameter names correspond to variable names given in SAS code from appendix 2.

Parameter	Estimate	Std. Error	Chi-Square	p
Intercept	-7.335	0.943	60.489	<0.001
all	-0.049	0.291	0.029	0.866
cmv	-0.140	0.241	0.339	0.560
male	0.141	0.245	0.330	0.566
age	0.045	0.060	0.551	0.458
gvhd	0.986	0.676	2.123	0.145
platnorm	-1.107	0.420	6.944	0.008
daysnoplatnorm	0.000	0.001	0.110	0.740
relapse	3.117	0.283	121.603	<0.001
daysnorelapse	0.000	0.001	0.059	0.808
agesq	0.000	0.001	0.107	0.744
day	-0.003	0.004	0.666	0.415
daysq	0.000	0.000	0.000	0.990
daycu	0.000	0.000	0.027	0.869
wait	0.009	0.011	0.596	0.440
gvhd*day	-0.002	0.005	0.196	0.658
gvhd*daysq	0.000	0.000	0.309	0.579
gvhd*daycu	0.000	0.000	0.396	0.529