“Completing the life course." Estimating the impact of hypothetical population-policies on cohort life expectancy with the g-formula: a worked example using hypertension and mortality in South Africa

EPC Extended Abstract

Maarten J. Bijlsma[1], Alpha Oumar Diallo[2], Nikkil Sudharsanan[3]

[1]Laboratory of Population Health, Max Planck Institute for Demographic Research
[2]Department of Epidemiology, University of North Carolina at Chapel Hill
[3]Heidelberg Institute of Global Health, Heidelberg University

Abstract

A key part of health-decision making is estimating how proposed interventions will affect the mortality of future cohorts. These types of questions are typically answered using modeling studies that draw estimates from multiple sources to simulate the life course of individuals. Although these studies are very common and influential, they have three important limitations, especially for developing countries: (1) by drawing estimates from multiple sources, they assume that the effect of an intervention on mortality from one population can be transported to other countries and populations; (2) they generally require comprehensive mortality registration information, which is often unattainable in developing countries; and (3) they make strong stationarity assumptions and assume that period mortality and health conditions accurately represent the dynamic experience of an aging cohort of individuals. In this paper, we propose an alternative approach that overcomes some of these limitations using longitudinal survey data and the parametric g-formula -- an epidemiological dynamic causal inference model. Specifically, we first estimate mortality and risk factor transitions as a function of age and potential confounders from a real cohort of individuals. We then use this information to project the covariate trajectories of the cohort beyond ages observed in the data and then complete their life course by estimating future mortality as a function of these covariate trajectories. This allows us to estimate the impact of population-policies on cohort life expectancies without having to transport estimates from one context to another while also relaxing stationarity assumptions by incorporating projected cohort covariate trajectories into future predictions of mortality. We describe and demonstrate this approach using a worked example of blood pressure control in South Africa.

Introduction

A key part of health-decision making is estimating and comparing how different interventions might affect the health of populations. For researchers and policy makers seeking to improve longevity, this translates into knowing how proposed interventions will affect the mortality of cohorts that will ultimately receive the interventions. These types of questions are extremely challenging to answer with traditional policy evaluation designs such as randomized control trials (RCTs) because they would require following cohorts until every member dies and somehow ensuring compliance to the intervention across this entire period. For this reason, questions on the cohort mortality impact of interventions are typically answered using modeling studies that simulate the life course of individuals (1,2). Researchers usually simulate several policy scenarios and compare cohort mortality across scenarios; for example, researchers may compare a policy that is administered to all individuals versus one administered to high-risk individuals (3). This information is then often linked to cost or effort data to determine which policy scenario is most cost-effective or feasible (4,5).

Health policy models are generally built using the following procedure (1,6,7). First an initial cohort of individuals at some starting age is drawn. These individuals are then aged forward based on mortality rates drawn from national death statistics and information on the age-specific prevalence of the intervention target (e.g. blood pressure [BP] if the policy was aimed at estimating the effect of BP treatments). This cohort usually forms the "natural course" comparison group since the main intervention target (in our example BP) is not changed in any way. Researchers then form an intervention cohort where the policy of interest is applied. The effect of the policy on mortality in this cohort is simulated based on effect sizes drawn from clinical trials or long running cohort data.

For example, suppose we were interested in the effect of treating systolic BP down to 125 mmHg among those aged 30 and above on cohort survival. We would begin by generating a

population starting at age 30. Next we would use national death statistics to create two sources of mortality: background mortality that is not affected by BP and "BP-amenable" mortality. We would also draw information on the age-specific prevalence of BP from a population survey. We would then simulate these 30-year olds forward using the age-specific mortality rates and the age-specific information on BP prevalence. For our intervention cohort we would repeat this process, this time making sure that individuals have a systolic BP that never exceeds 125 mmHg. At every age, we would estimate the effect of this BP reduction on survival by reducing BP-amenable mortality by an amount based on clinical trial data on the effect of BP reductions on mortality. This cohort would then be survived forward like the natural course cohort and we would compare cohort life expectancy between the natural course and intervention scenarios.

While these types of studies are extremely common and very influential, they have three important limitations -- two of which are especially pronounced for developing country contexts. First, these studies assume that the effect of the policy on mortality drawn from clinical trials and cohort studies from one context accurately represent the mortality reductions that would occur at the population level in other contexts. This assumption is known as transportability and is particularly strong when clinical trial or cohort data from high-income countries is used to evaluate policy scenarios in low- and middle-income countries (8). Second, estimating the effect of the policy on mortality for the simulated cohorts usually involves partitioning mortality into background and intervention-amenable mortality. This requires cause-of-death data or at a minimum comprehensive mortality registration information, both of which may be unattainable in developing countries (9). Lastly, these models make strong stationarity assumptions: information on mortality and risk-factor progressions over age are drawn from period data sources, like national life tables and population surveys, but then used to simulate cohort life courses. This assumes that mortality and risk factor

dynamics have been stationary over time and that the cross-sectional pattern accurately represents the dynamic experience of a real aging cohort of individuals (10).

A small literature has emerged that proposes using the parametric g-formula, an epidemiological simulation method (11), as an alternative to standard health policy modeling to overcome some of these limitations (12). First, the parametric g-formula approach involves estimating the effect of the intervention on mortality using data from the target population themselves. This removes the need to transport an effect from another context.[1] Second, the g-formula approach bypasses the need for vital registration data by using micro-level data with mortality follow-up information. Lastly, the g-formula approach is based on cohort, not cross-sectional, data and thus does not involve making stationarity assumptions. The fundamental limitation to the g-formula approach as it is often applied to mortality questions (13), however, is that it is limited in years of follow up by the number of survey waves in the data. Therefore, in order to be an alternative to health policy modeling, the researcher would need cohort data with regular and repeated measurements that span the entire life course of surveyed individuals. This is a requirement that cannot currently be realistically met, to the best of our knowledge, with data from any context.

The g-formula approach addresses the transportability and stationarity issues of typical health policy modeling but has unrealistic data assumptions when cohort life expectancies are the ultimate outcome of interest. In this paper, we propose and demonstrate an extension of the parametric g-formula that addresses this key data limitation by combining it with aspects of traditional health policy and demographic simulation approaches. Specifically, we first estimate mortality and risk factor transitions as a function of age from a real cohort of individuals. We then use this information to project the covariate trajectories of the cohort beyond ages observed in the data and then complete

[1] This effect, however, is observationally estimated, leading to a tradeoff between bias due to unobserved confounding and bias due to transportability. We discuss this issue in greater detail later in the paper.

their life course by estimating future mortality as a function of these covariate trajectories. This allows us to combine the advantages of health policy modeling (can cover the entire life course of a cohort under both natural course and intervention scenarios) with the advantages of the g-formula (we can avoid the transportability assumption and relax the stationarity assumption by incorporating information on cohort covariate trajectories into future predictions of mortality). In the following sections, we describe this approach and provide a step-by-step worked example that estimates the effect of population-policies to improve blood pressure control on cohort life expectancies using data from South Africa.

Concepts and intuition behind the approach

We describe our approach through the following motivating example: How does cohort life expectancy for 30-year olds in South Africa change if there is a population-policy to keep their systolic BP under control (≤ 125 mmHg) for the rest of their lives?

Target trial

To begin answering this question, it is useful to describe the hypothetical randomized control trial, that if possible to run, would provide us with an answer to our motivating question (this is sometimes known as the "target trial") (14). First, we would select a population-representative cohort of South African 30-year olds. We would then randomly split this cohort into an intervention and control group. For the intervention group, we would, through a combination of medicines and lifestyle changes, ensure that systolic BP never exceeds 125 mmHg for the rest of the cohort members' lives. For the control group, we would let them age as is - this means that some individuals may independently take up medicine and achieve full control, others may achieve partial control or control for only a part of their lives, and others may never be able to control their BP (this is sometimes referred to as the "current best practices" cohort). We would then follow these two cohorts until every individual dies and estimate the effect of the policy on cohort life expectancy by a simple comparison of the average age of death between the intervention and control cohorts.

Real data and the parametric g-formula

In reality, this type of trial is nearly impossible to run. Our goal, therefore, is to try and use longitudinal survey data to mimic the target trial as closely as possible. Suppose, for example, we had data on adults from a longitudinal population-representative survey collected in South Africa. These types of data sources are common but present a fundamental challenge for health policy modeling:

we only have data on any given individual for the limited window of time that the survey has been running.

Our solution to this issue is to extrapolate the parametric g-formula, a method for estimating dynamic causal effects in epidemiology, to simulate our target trial intervention and control cohorts. The core principle behind the parametric g-formula is to first estimate relationships from empirical and data then simulate the life course of cohorts based on the longitudinal relationships found in the data. Typically, however, this is only done for the number of waves present in the actual survey (13,15). What we propose is to project the cohort beyond the years found in the data. For example, suppose our survey data covered a 10 year period. Therefore, for a 30-year old, the data only tells us what would happen until that individual is 40. To complete this individual's life course, we project the empirical relationships found in the data beyond age 40, predicting how that individual's covariate and mortality trajectories would evolve if they were observed beyond age 40. This approach is similar in some ways to the synthetic cohort approaches common in demography and health policy but is fundamentally different in one crucial way. Rather than just assuming that a 30-year old in the data will have the mortality experience at age 40 of a 40-year old in the data (the stationarity assumption), we actually try and estimate how that 30-year old would look when they are 40 in terms of their covariates, and then predict what mortality at that age would be given those new covariate values.

This approach allows us to use longitudinal survey data to recreate the target trial we described previously in two steps. We do this by applying the parametric g-formula in two ways. First, we simulate a cohort that follow the observed aging patterns in the empirical data. Since we are not simulating any type of intervention, this cohort forms the control group of the target trial. Next, we simulate a second cohort where, rather than allowing blood pressure to evolve among individuals in the way it does in the data, we constrain BP to never exceed 125 mmHg. By holding BP at a controlled

level, we also end up affecting the mortality rates experienced by the cohort. This new cohort, because they have "aged" under this BP restriction, form our treatment group. Lastly, just like our target trial, we then evaluate the impact of the intervention on cohort life expectancy by comparing the average age at death across the control and intervention cohorts.

In the following section, we provide a step-by-step worked example of this process using data from the South African National Income Dynamics Study.

Worked example

Data

Data for this illustrative example are from the 2008-2017 of the South African National Income Dynamics Survey (NIDS) (16). The NIDS is a nationally representative survey of individuals of all ages and contains extensive demographic, economic, and health information. To keep our example simple, we consider just 7 variables from the NIDS: age (in years), sex (male/female), measured mean systolic blood pressure (based on the average of 2 measurements taken with an electronic blood pressure monitor), measured body mass index (based on measured height and weight), current smoking status (0/1), schooling (no schooling, grade school, higher education degree), and whether an individual died between survey waves (0/1).[2] To be consistent with our target trial, we focus on adults ages 30 and above and for the sake of the worked example limit our sample to individuals who have non-missing data for every wave of data that they are alive in.

Before conducting our analyses, we first convert the data to a person-wave format with observations for every wave that an individual is observed in. For each person-wave, we classify the mortality variable as 0 if the individual survived to the subsequent wave or 1 if they died before the subsequent wave of data was collected. Our total sample consists of 4,724 observations corresponding to 18,909 person-waves of data.

[2] Schooling and sex are the only time-invariant variables we consider.

Estimating the relationships in the data

Our first step is to use parametric models to estimate how mortality evolves over age. This model forms the basis of the simulation by providing a way of estimating the probability of mortality at any given age. We may be tempted to fit the following simple model (shown for just one sex):

$$logit(E[D|A]) = \alpha_0 + (\alpha_1 * A)$$

where D is the 0/1 indicator of whether an individual died between waves and A is age. This is problematic, however, because from a causal perspective, this model assumes that there are no confounders of the age-mortality relationship (**Figure 1**):

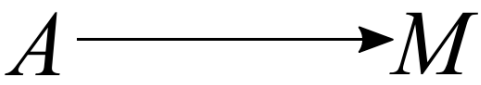


Figure 1. Directed Acyclic Graph (DAG) of the effect of age (A) on mortality (M).

The problem with this assumption is that since our data are from a specific period (2008-2017), individuals in the data at different ages come from multiple birth cohorts. Therefore, the model's estimate of what mortality for a 30-year old would be when they reach age 40 is effectively estimated from the mortality experience of 40-year olds in that same period (similar to the synthetic cohort approach). If there were no differences across birth cohorts in characteristics relevant to mortality (the stationarity assumption), this would not actually be a problem, and could be represented by the following causal diagram (**Figure 2**):

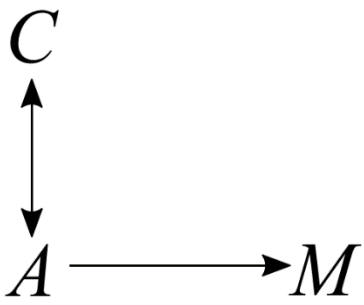


Figure 2. Directed Acyclic Graph (DAG) of the effect of age (A) on mortality (M), including birth cohort (C).

Here C is a set of indicators for birth cohort. The causal diagram above appears to assume that differences in characteristics across cohort do not affect mortality, and therefore that cohort is not a common cause of both age and mortality and hence does not bias this relationship. However, in reality, there are likely to be several differences across cohorts that are also related to mortality. For our simplified example, we will assume that there are just four mortality-relevant characteristics that vary across cohorts, systolic BP (B), BMI (represented by W for weight), tobacco use (T), and schooling (S) (**Figure 3**):

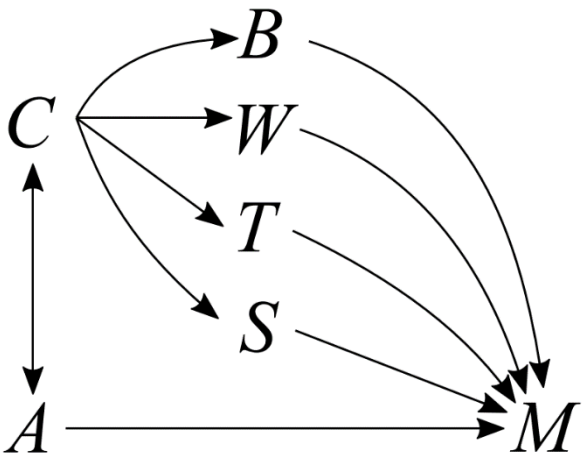


Figure 3. Directed Acyclic Graph (DAG) of the effects of age (A) and birth cohort (C) on mortality, via characteristics of birth cohort, i.e. systolic blood pressure (B), BMI (W), tobacco use (T) and schooling (S).

For now we will also assume that there is no direct effect of age on any of these variables (an assumption we will soon remove). This diagram reveals that the relationship between age and mortality estimated from the previous model is confounded by differences in schooling, tobacco use, BMI, and BP across ages that are due to the fact that individuals at different ages in the data come from different birth cohorts. One way to address this confounding is to include these variables in our model[3]:

$$logit(E[D|A,B,W,T,S]) = \alpha_0 + (\alpha_1 * A) + (\alpha_2 * B) + (\alpha_3 * W) + (\alpha_4 * T) + (\alpha_5 * S)$$

The main issue with this model and causal structure, however, is that we are assuming that there is no effect of age on any of the intermediary variables. In reality, there is a relationship between age and tobacco use, systolic BP, and BMI that is not just driven by differences across cohorts (**Figure 4** shows this relationship for systolic BP and BMI).

[3] A large literature in demography, sociology, epidemiology and statistics has worked on age-period-cohort models, with the fundamental problem that it is not possible to simultaneously estimate the role of all three since Age = Period - Cohort. One alternative that has been proposed is to examine age, period, or cohort effects not through the use of indicator variables for these variables but by modeling proxies for at least one of the age, period or cohort variables (17–19). Our approach effectively takes this strategy, modeling the "descendants" of cohort, such as smoking and obesity. In our case, this allows a transparent solution to the APC problem, but additional it allows us to directly vary these characteristics to form different counterfactual estimates - which is needed for our approach (described later on) and would not be possible if these effects were captured as part of a cohort indicator.

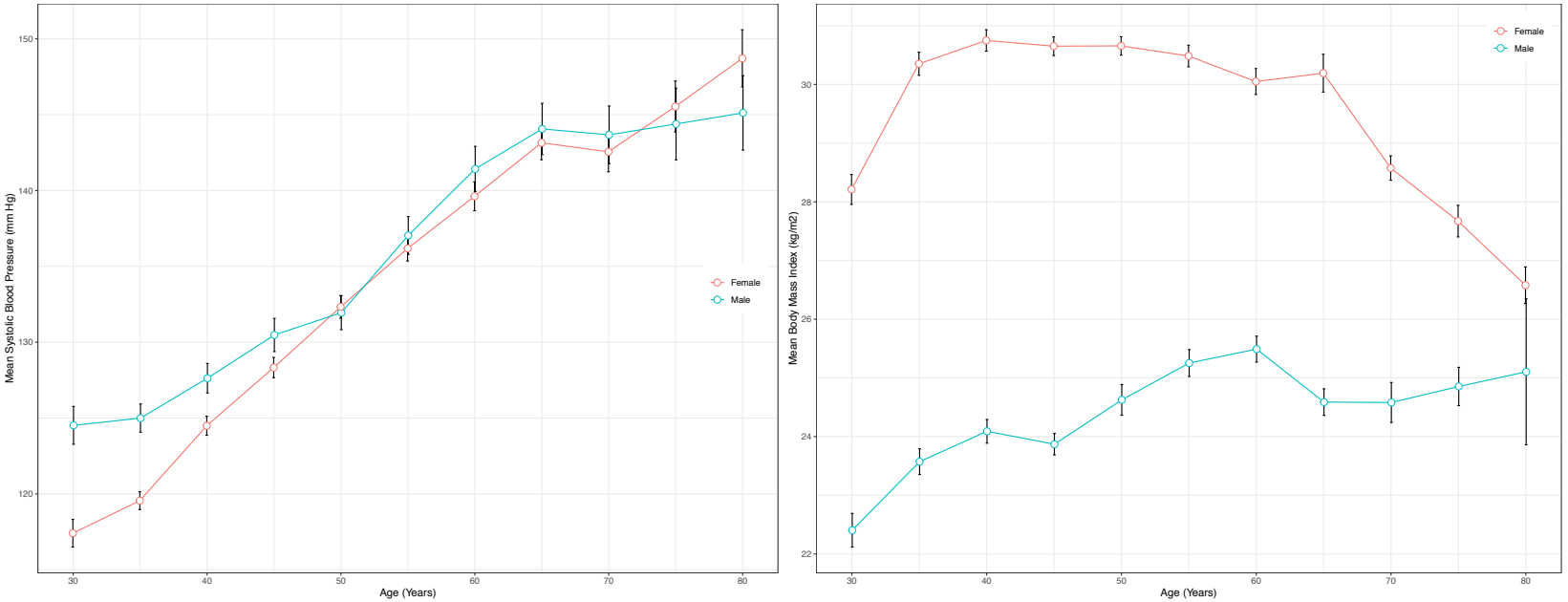


Figure 4. Age-patterns of systolic blood pressure and BMI by sex, adults ages 30+, National Income Dynamics Study, 2008-2017.

Therefore, a more realistic causal structure might be (**Figure 5**):

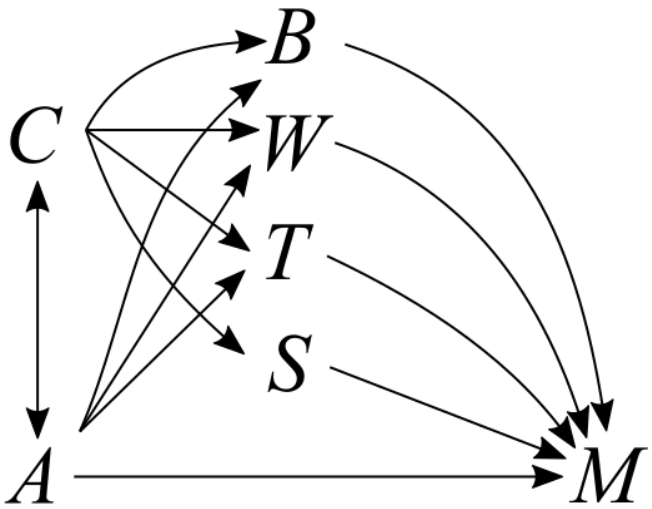


Figure 5. Directed Acyclic Graph (DAG) of the effects of age (A) and birth cohort (C) on mortality, via characteristics of age and birth cohort, i.e. systolic blood pressure (B), BMI (W), tobacco use (T) and schooling (S).

We now have an issue. Tobacco use, BMI, and systolic BP are now both confounders and mediators of the age - mortality relationship. This means that a regression that simply includes these variables as controls would result in incorrect estimates of mortality rates by age for any given birth cohort since we have "controlled away" the part of the direct effect of age on mortality that runs through tobacco use, BMI, and blood pressure. Said another way, the model would predict how mortality changes over age, net of the effect of age on tobacco use, BMI, and systolic BP, and thus would not represent how the actual cohort ages because in reality, these three characteristics do change over age. The parametric g-formula provides a solution to this problem by not just modeling the relationship between mortality and age, but also age and every variable it may affect (the time or age-varying variables)[4,5]:

[4] Since our data come from survey waves separated by two years, the mortality model is predicting the two-year probability of death while the other models are predicting the two-year change in each variable.

[5] We have included lagged terms in this model to indicate that the value of any age-varying variable at a given age is also related to its value at the previous age.

$$logit(E[D_{a+2}|A_a, B_a, W_a, T_a, S]) = \alpha_0 + (\alpha_1 * A_a) + (\alpha_2 * B_a) + (\alpha_3 * W_a) + (\alpha_4 * T_a) + (\alpha_5 * S)$$

$$E[B_{a+2}|A_a, B_a, W_a, T_a, S] = \beta_0 + (\beta_1 * A_a) + (\beta_2 * B_a) + (\beta_3 * W_a) + (\beta_4 * T_a) + (\beta_5 * S)$$

$$E[W_{a+2}|A_a, B_a, W_a, T_a, S] = \gamma_0 + (\gamma_1 * A_a) + (\gamma_2 * B_a) + (\gamma_3 * W_a) + (\gamma_4 * T_a) + (\gamma_5 * S)$$

$$logit(E[T_{a+2}|A_a, B_a, W_a, T_a, S]) = \delta_0 + (\delta_1 * A_a) + (\delta_2 * B_a) + (\delta_3 * W_a) + (\delta_4 * T_a) + (\delta_5 * S)$$

This set of models can be represented by our final causal diagram (**Figure 6**):

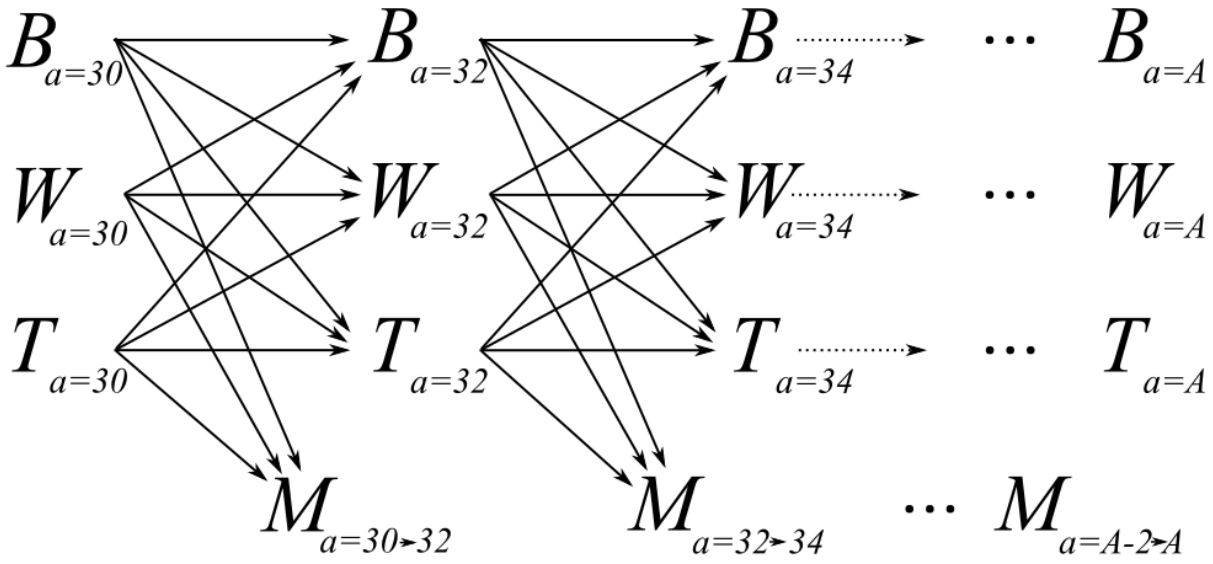


Figure 6. Cross-lagged DAG showing the relationship between the time-varying variables. Mortality (M) is measured between the waves. Age and Schooling are not shown for simplicity: they affect all time-varying variables. The time varying variables are systolic blood pressure (B), BMI (W), and tobacco use (T).

Under this system, we estimate mortality not based on a single model but on a set of models that feed into one another in an order that preserves the relationship between age and the intermediary variables while also controlling for differences in these variables that are due to cohort effects. We are now ready for our first step:

Step 1: Estimate relationships in the data based on the set of models corresponding to the causal structure in Figure 6 (regression results are presented in the first panel of Figure 7).

Creating the natural course control cohort

After fitting models to capture the relationships in the empirical data, the next step is to simulate the control cohort of the target trial. Our goal in this step is to estimate how long the current cohort of 30-year olds will live on average under the natural course or status quo scenario where no new BP intervention is introduced. We describe this approach algorithmically here, bringing us to step 2:

Step 2: Simulate the natural course cohort under the causal structure estimated as part of step 1.

1. Create a dataset with a large number of pseudo-30-year olds drawn from the original data (for this example, we will draw 3000 individuals). By drawing these individuals from the original data, our pseudo-cohort has the baseline covariate distribution -- and covariance between these covariates -- found in the empirical data.
2. Now, for each 30-year old, estimate their probability of surviving to age 32 (since the data correspond to 2-year survival probabilities) given their covariate values by inputting their covariate values into the mortality regression equation.
3. Draw a 0/1 value from this probability for each individual (based on a binomial distribution) to determine which individuals survived and which did not.
4. For individuals that died (drew a D = 1), stop the simulation at this point.
5. For individuals that survived (drew a D=0):
    a. Deterministically update age to 32 and carry forward the values of the time/age-invariant covariates.
    b. Estimate the distributional parameters (mean and standard deviation for normal variables and probability for binomial variables) of each of the time/age-varying covariates for each individual when they are 32 by plugging their covariate information into the tobacco, BMI, and sys BP regression equations.

c. Predict values for the time-varying covariates for each individual when they are 32 by drawing from distributions with the parameters estimated in the previous step. We now have a fully updated set of values for this cohort at age 32.

6. Repeat this process until every member of the cohort "dies."

We demonstrate this process for the age 30 to 32 transition for a hypothetical member of the natural course cohort in **Figure 7**.

Step 1: Regression estimates

$$logit(E[D_{a+2}|A_a, BP_a, BMI_a, SMK_a, SomeSCH, HighSCH]) = \widehat{-5.5} + (\widehat{0.048} * A_a) + (\widehat{0.0002} * BP_a) + (\widehat{-0.004} * BMI_a) + (\widehat{0.43} + SMK_a) + (\widehat{-0.15} * SomeSCH) + (\widehat{-0.33} * HighSCH)$$

$$E[BP_{a+2}|A_a, BP_a, BMI_a, SMK_a, SomeSCH, HighSCH] = \widehat{54.9} + (\widehat{0.48} * A_a) + (\widehat{0.33} * BP_a) + (\widehat{0.26} * BMI_a) + (\widehat{1.56} * SMK_a) + (\widehat{1.61} * SomeSCH) + (\widehat{-0.25} * HighSCH)$$

$$E[BMI_{a+2}|A_a, BP_a, BMI_a, SMK_a, SomeSCH, HighSCH] = \widehat{13.0} + (\widehat{-0.006} * A_a) + (\widehat{0.004} * BP_a) + (\widehat{0.54} * BMI_a) + (\widehat{-0.30} + SMK_a) + (\widehat{0.92} * SomeSCH) + (\widehat{1.78} * HighSCH)$$

$$logit(E[SMK_{a+2}|A_a, BP_a, BMI_a, SMK_a, SomeSCH, HighSCH]) = \widehat{-2.3} + (\widehat{-0.011} * A_a) + (\widehat{0.0001} * BP_a) + (\widehat{-0.012} * BMI_a) + (\widehat{2.28} + SMK_a) + (\widehat{0.16} * SomeSCH) + (\widehat{-0.17} * HighSCH)$$

Step 2: Simulate the natural course cohort

Age 30 values (example male cohort member)

$BP_{30} = 110$
$BMI_{30} = 22$
$SMK_{30} = 1$
$SomeSCH = 0, HighSCH = 1$

Estimate probability of death before age 32

$$\hat{E}[D_{32}|A_{30} = 30, BP_{30} = 110, BMI_{30} = 22, SMK_{30} = 1, SomeSCH = 0, HighSCH = 1]$$

$$= \frac{e^{-5.5}e^{(0.048*30)}e^{(0.0002*110)}e^{(-0.004*22)}e^{(0.43*1)}e^{(-0.33*1)}}{1 + e^{-5.5}e^{(0.048*30)}e^{(0.0002*110)}e^{(-0.004*22)}e^{(0.43*1)}e^{(-0.33*1)}} = 0.018$$

Do they die before 32?

binomial.draw(probability = 0.018, draws = 1) -> 0

Estimate distribution parameter values of time-varying covariate values for age 32

$$\hat{E}[BP_{32}|A_{30} = 30, BP_{30} = 110, BMI_{30} = 22, SMK_{30} = 1, SomeSCH = 0, HighSCH = 1] = 54.9 + (0.48 * 30) + (0.33 * 110) + (0.26 * 22) + (1.56 * 1) + (-0.25 * 1) = 112.6$$

$$\widehat{SD}(BP_{32}) = \boldsymbol{sd}(\boldsymbol{residuals}(\text{BP Model})) = 6.2$$

$$\hat{E}[BMI_{32}|A_{30} = 30, BP_{30} = 110, BMI_{30} = 22, SMK_{30} = 1, SomeSCH = 0, HighSCH = 1] = 13.0 + (-0.006 * 30) + (0.004 * 110) + (0.54 * 22) + (-0.30 * 1) + (1.78 * 1) = 26.6$$

$$\widehat{SD}(BMI_{32}) = \boldsymbol{sd}(\boldsymbol{residuals}(\text{BMI Model})) = 7.7$$

$$\hat{E}[SMK_{32}|A_{30} = 30, BP_{30} = 110, BMI_{30} = 22, SMK_{30} = 1, SomeSCH = 0, HighSCH = 1] = \frac{e^{-2.3}e^{(-0.011*30)}e^{(0.0001*110)}e^{(-0.012*22)}e^{(2.28*1)}e^{(-0.17*1)}}{1 + e^{-2.3}e^{(-0.011*30)}e^{(0.0001*110)}e^{(-0.012*22)}e^{(2.28*1)}e^{(-0.17*1)}} = 0.32$$

Simulate and set covariate values for age 32

BP_{32} = **normal.draw**(mean = 112.6, sd = 6.2) -> 106.6
BMI_{32} = **normal.draw**(mean = 26.6, sd = 7.7) -> 25.8
SMK_{32} = **binomial.draw**(probability = 0.32, draws = 1) -> 0
$SomeSCH = 0, HighSCH = 1$

Figure 7: Flowchart demonstrating the parametric g-formula simulation process for the 30-32 transition for one hypothetical male cohort member.

Creating the intervention cohort

The next step is to simulate the intervention cohort of the target trial. Our aim is to estimate how long the current cohort of 30-year olds will live on average in a world where their blood pressure is controlled such that it never exceeds 125 mmHg. This process is similar to the natural course simulation with the main difference being that we need to constrain blood pressure to never increase above 125 mmHg.

Step 3: Simulate the intervention cohort under the causal structure (Figure 6) AND the specified intervention scenario

1. Create a dataset with a large number of pseudo-30-year olds drawn from the original data (for this example, we will draw 3000 individuals). By drawing these individuals from the original data, our pseudo-cohort has the baseline covariate distribution -- and covariance between these covariates -- found in the empirical data.
2. Before predicting survival, set BP for all individuals above 125 mmHg down to 125.
3. Now, for each 30-year old, estimate their probability of surviving to age 32 given their covariate values by inputting their covariate values into the mortality regression equation.
4. Draw a 0/1 value from this probability for each individual to determine which individuals survived and which did not.
5. For individuals that died (drew a D = 1), stop the simulation at this point.
6. For individuals that survived (drew a D=0):
    a. Deterministically update age to 32 and carry forward the values of the time/age-invariant covariates.

b. Estimate the distributional parameters of each of the time/age-varying covariates for each individual when they are 32 by plugging their covariate information into the tobacco, BMI, and sys BP regression equations.
c. Predict values for the time-varying covariates for each individual when they are 32 by drawing from distributions with the parameters estimated in the previous step.
d. As before, replace BP for all individuals with an updated BP above 125 mmHg down to 125 mmHg. We now have a fully updated set of values for this cohort at age 32.

7. We then repeat this process until every member of the cohort "dies."

Estimating the impact of the population-policy on cohort survival

At the end of this procedure we have effectively simulated our target trial. The natural course or control cohort of the target trial is represented by the cohort of individuals for whom we did not manipulate their BP and the intervention cohort are represented by the cohort of individuals that we held at a maximum BP of 125 mmHg for their entire life course. This leads to the last step:

Step 4: Compare the natural course and intervention cohorts.

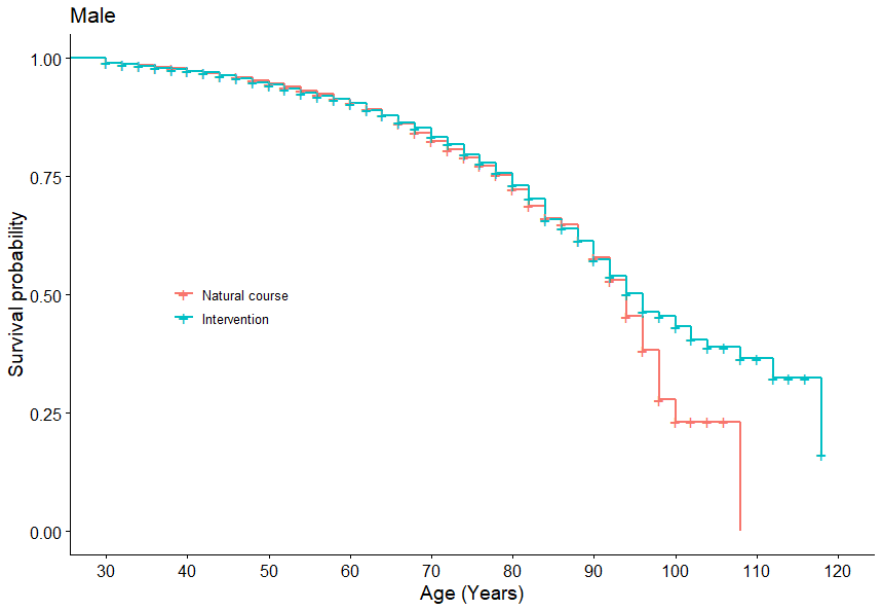
To compare the population-level effect of this intervention on cohort life expectancy at age 30, we now directly compare the mean age at death across the two pseudo populations (**Table 1**).

Table 1: Comparison of g-formula estimated cohort life expectancy for South African 30-year olds between the natural course and intervention (systolic BP never exceeds 125 mmHg) scenarios.

	Natural course	Intervention
Male	53.9	53.5
Female	65.3	64.1

In this example, we do not find evidence that controlling blood pressure across the life course does not have a beneficial impact on cohort life expectancy for 30-year olds in South Africa (these results are from preliminary analyses and should be interpreted with caution).

Beyond just the cohort life expectancy, we can also compare the survival curves between the two cohorts to determine the impact of the intervention on death distributions (**Figure 8**):



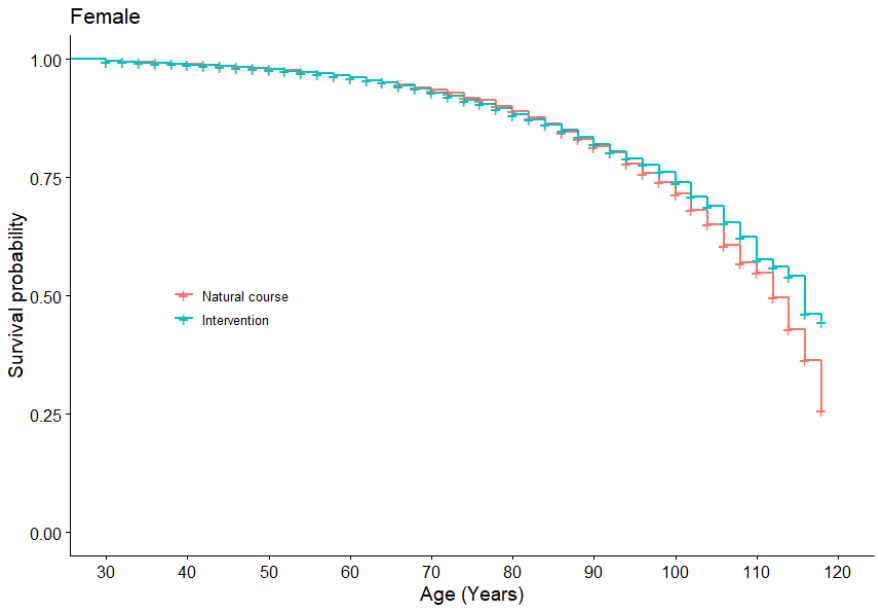


Figure 8: Comparison of Kaplan-Meyer survival curves for South African 30-year olds between the natural course and intervention (systolic BP never exceeds 125 mmHg).

The survival curves reveal that there may be some benefits to BP control in the advanced ages that is not being reflected in the measures of life expectancy.

Next Steps

We plan to complete the following additional steps over the next several months

1. We will include a section on model diagnostics, focusing on e.g. making sure the predicted covariate trajectories stay within realistic ranges and the extent to which the model correctly captures the observed data.
2. We will include a discussion on the estimation of confidence intervals and correcting for the Monte Carlo error introduced by drawing discrete values from distributions.
3. We are working on and will include a detailed section on causal inference assumptions as they relate to this approach. Specifically, we will discuss the so-called consistency and

exchangeability assumptions, have a deeper discussion on transportability, and how this approach should be interpreted in light of these assumptions.

References

1. Jit M, Brisson M, Portnoy A, Hutubessy R. Cost-effectiveness of female human papillomavirus vaccination in 179 countries: a PRIME modelling study. Lancet Glob Health. 2014;2(7):e406–e414.

2. Capewell S, Ford ES, Croft JB, Critchley JA, Greenlund KJ, Labarthe DR. Cardiovascular risk factor trends and options for reducing future coronary heart disease mortality in the United States of America. Bull World Health Organ. 2010 Feb 1;88(2):120–30.

3. Basu S, Shankar V, Yudkin JS. Comparative effectiveness and cost-effectiveness of treat-to-target versus benefit-based tailored treatment of type 2 diabetes in low-income and middle-income countries: a modelling analysis. Lancet Diabetes Endocrinol. 2016;4(11):922–932.

4. Pandya A, Doran T, Zhu J, Walker S, Arntson E, Ryan AM. Modelling the cost-effectiveness of pay-for-performance in primary care in the UK. BMC Med. 2018;16(1):135.

5. Adam T, Murray C. Making choices in health: WHO guide to cost-effectiveness analysis. Vol. 1. World Health Organization; 2003.

6. Ford ES, Ajani UA, Croft JB, Critchley JA, Labarthe DR, Kottke TE, et al. Explaining the decrease in US deaths from coronary disease, 1980–2000. N Engl J Med. 2007;356(23):2388–2398.

7. Pandya A, Sy S, Cho S, Alam S, Weinstein MC, Gaziano TA. Validation of a cardiovascular disease policy microsimulation model using both survival and receiver operating characteristic curves. Med Decis Making. 2017;37(7):802–814.

8. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiol Camb Mass. 2011;22(3):368.

9. Mikkelsen L, Phillips DE, AbouZahr C, Setel PW, De Savigny D, Lozano R, et al. A global assessment of civil registration and vital statistics systems: monitoring data quality and progress. The Lancet. 2015;386(10001):1395–1406.

10. Guillot M. Period versus cohort life expectancy. In: International handbook of adult mortality. Springer; 2011. p. 533–549.

11. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model. 1986;7(9–12):1393–1512.

12. Murray EJ, Robins JM, Seage GR, Freedberg KA, Hernán MA. A comparison of agent-based models and the parametric g-formula for causal inference. Am J Epidemiol. 2017;186(2):131–142.

13. Keil AP, Edwards JK, Richardson DR, Naimi AI, Cole SR. The parametric G-formula for time-to-event data: towards intuition with a worked example. Epidemiol Camb Mass. 2014;25(6):889.

14. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. Am J Epidemiol. 2016;183(8):758–764.

15. 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):1599–1611.

16. Southern Africa Labour and Development Research Unit. National Income Dynamics Study (NIDS) Wave 1, 2008 [dataset]. Cape Town; 2016.

17. O'Brien RM. Age period cohort characteristic models. Soc Sci Res. 2000;29(1):123–139.

18. Winship C, Harding DJ. A mechanism-based approach to the identification of age–period–cohort models. Sociol Methods Res. 2008;36(3):362–401.

19. Bijlsma MJ, Daniel RM, Janssen F, De Stavola BL. An assessment and extension of the mechanism-based approach to the identification of age-period-cohort models. Demography. 2017;54(2):721–743.