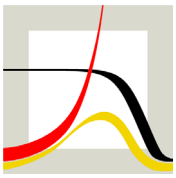


Max-Planck-Institut für demografische Forschung

Max Planck Institute for Demographic Research

Konrad-Zuse-Strasse 1 ▪ D-18057 Rostock ▪ Germany ▪ Tel +49 (0) 3 81 20 81 - 0 ▪ Fax +49 (0) 3 81 20 81 - 202 ▪ www.demogr.mpg.de

MPIDR Working Paper WP 2019-004 I February 2019

A Generalized Counterfactual Approach to Decomposing Differences Between Populations

Nikkil Sudharsanan
Maarten J. Bijlsma I bijlsma@demogr.mpg.de

This working paper has been approved for release by: Kieron Barclay (barclay@demogr.mpg.de), Deputy Head of the Laboratory of Population Health.

A Generalized Counterfactual Approach to Decomposing Differences Between Populations

Nikkil Sudharsanan[a,b] and Maarten J. Bijlsma[c]

[a]Heidelberg Institute of Global Health, Heidelberg University
[b]Harvard T.H. Chan School of Public Health, Harvard University
[c]Laboratory of Population Health, Max Planck Institute for Demographic Research

Abstract

One central aim of the population sciences is to understand why one population has different levels of health and well-being compared to another. Various methods, such as the Oaxaca-Blinder and Kitagawa decompositions, have been used to decompose population-differences in a wide range of outcomes. We provide a way of implementing an alternative decomposition method that, under certain assumptions, adds a causal interpretation to the decomposition by building upon counterfactual-driven estimation methods. In addition, the approach has the advantage of flexibility to accommodate different types of outcome and explanatory variables and any population contrast. By using Monte Carlo methods, our approach does not rely on closed-form approximate solutions and can be applied to any parametric model without having to derive any decomposition equations. We demonstrate our approach through two motivating examples using data from the Mexican Health and Aging Study and the 1970 British Birth Cohort Study. The first example uses a cross-sectional binary outcome (disability), a contrast of prevalence rates, and considers a binary mediator (stroke), while the second example uses a count outcome (age at first birth), a contrast of median ages, and considers a count mediator (women's own years of education). Together, our two examples outline how to implement a very generalized decomposition procedure that is theoretically grounded in counterfactual theory but still easy to apply to a wide range of situations. We provide example R-code and an R-function [package in development].

Keywords: decomposition, causal inference, Monte Carlo, parametric g-formula, population models

Introduction

One central aim of the population sciences is to understand why one population has different levels of health and well-being compared to another. Recent examples of this type of research include understanding why African Americans have worse health compared to white Americans (Geruso 2012; Kittner et al. 1990), why the United States has lower life expectancy compared to other high-income countries (Ho 2013), why poorer individuals in Finland have higher mortality compared to more affluent individuals (Martikainen et al. 2014), and why the southern American states have higher rates of cardiovascular disease compared to other parts of the country (Steckel and Senney 2015). By identifying the sources of differences across populations, these studies provide an important first step for determining what can be done to reduce disparities.

Different disciplines have developed various methods to answer this wide array of questions. For example, in economics and demography, the Oaxaca-Blinder (Blinder 1973; Oaxaca 1973), Kitagawa (Kitagawa 1955), Arriaga (Arriaga 1984), and related decompositions (Andreev, Shkolnikov, and Begun 2002; Chevan and Sutherland 2009; Gupta 1978; Horiuchi, Wilmoth, and Pletcher 2008) have been used extensively to decompose population-differences in various outcomes. While computationally simple, these decompositions often do not have a clear counterfactual interpretation or causal structure. This creates ambiguity about the meaning of the decomposition results and thus inhibits their value for identifying the causes of disparities (we demonstrate how traditional decompositions can produce misleading results in the presence of confounding variables in Appendix 1). Recent advances in epidemiology and psychology have provided an alternative to these decompositions (Jackson and VanderWeele 2018; Nandi, Glymour, and Subramanian 2014). Jackson and VanderWeele (JVW) (2018), develop a general decomposition that uses causal inference and counterfactual theory to provide decomposition estimates that are mapped to specific intervention scenarios and therefore have a clear counterfactual interpretation.

JVW demonstrate how this method can be used to decompose racial differences in wages, rates of incarceration, and educational test scores also prove that the Oaxaca-Blinder and Kitagawa decompositions are special cases of this counterfactual decomposition approach (Jackson and VanderWeele 2018).

There are two challenges that hinder the implementation of the counterfactual decomposition to answer questions that are common in demography and the population sciences. First, JVW only provide analytical solutions to their non-parametric equations for specific scenarios (decomposing differences in normally distributed or rare binomially distributed variables). Therefore, applying the counterfactual decomposition to other outcome variables common in the population sciences, such mortality (time-to-event), fertility (count), or disability (common binomial) requires that the researcher separately derive a new set of equations depending on the distribution of the outcome variable of interest, and in many cases, may require making approximations to obtain closed-form solutions. The second limitation is that the counterfactual decomposition is developed to decompose mean differences between populations, even though many important population outcomes, such as age-standardized prevalence rates, median survival times, and period life expectancy are not based around direct mean comparisons.

In this paper, we develop a Monte Carlo-based counterfactual decomposition that is easily applied to a wide range of questions common in demography and the population sciences. Our implementation does not require the analyst to derive decomposition equations and allows for flexible estimation with different types of outcome and explanatory variables. The Monte Carlo implementation also builds on the theory developed by JVW by allowing for comparisons in any population contrast, including the entire distribution of the outcome variable. In the first section of our paper, we introduce our approach. In the second and third section, we demonstrate this approach using two motivating examples: (1) 'why is disability higher among the lower educated in

Mexico compared to the more educated', and (2) 'why do women in the United Kingdom from lower socioeconomic backgrounds have children at younger ages compared to their more advantaged counterparts?' The first example uses a cross-sectional common binary outcome (disability), a contrast of prevalence rates, and considers a binary mediator (stroke), while the second example uses a count outcome (age at first birth), a contrast of median ages, and considers a count mediator (women's own years of education). Together, our two examples demonstrate how to implement a very generalized decomposition procedure that is theoretically grounded in counterfactual theory but still easy to apply to a wide range of situations.

A Counterfactual Approach to Decomposition

A common question in the population sciences is, "why is there a difference in outcome Y between individuals belonging to group A compared to individuals belonging to group B?" Specifically, researchers are often interested in the contribution of a cause of Y, M, to differences between groups A and B. We begin by specifying a concrete definition for the contribution of cause M to differences in outcome Y. Next, we provide a brief background on the counterfactual and potential outcomes theory that form the basis of the decomposition. We then discuss the estimation of the decomposition quantities through the use of parametric models.

To begin answering this question, our approach requires specifying a concrete definition for the "contribution" of cause M to differences in outcome Y. We adopt a counterfactual perspective and ask, "how large would differences in Y be if both groups A and B had an equal distribution of M?" Hence, the contribution of M is now defined as the change in Y before and after equalizing the distribution of M. We specifically examine the percent reduction and define the contribution as:

$$\text{Contribution} = 1 - \frac{\Delta Y_{counterfactual}}{\Delta Y_{observed}} \quad (1)$$

Note that we have not defined what exact contrast Δ represents; indeed, a major advantage of our approach is this generality because it allows for the easy comparison of multiple contrasts between A and B, such as mean or median differences in Y, or, if Y was an outcome like mortality, a difference of a function of Y such as life expectancy. Our examples will illustrate this flexibility in more detail.

Estimating the contribution (1) also requires specifying what distribution of M we want to equalize both groups to. A number of possible options exist. For example, if M is a discrete event, such as stroke, the probability of M could be set to any value between 0 and 100% for both populations. If M is normally distributed, we could equalize M to a normal distribution with any mean and standard deviation. When the relationship between Y and M is linear, the choice of this reference distribution does not affect the contribution; however, when Y and M have a non-linear relationship, the choice is non-arbitrary and different distributions can result in different estimates of the contribution of M (Andreev et al. 2002). For this reason, the choice of the reference level should be informed by substantive concerns (e.g. what level makes sense from a policy perspective?) and inferential concerns (e.g. certain values may be outside of the data). For example, a mean value of 0 for a discrete cause M implies no one in either population has the cause and may not be a realistically achievable scenario. A more appealing counterfactual might be to set the mean M in group B to be the same as the mean M in group A. We follow this approach for our exposition below.

To begin, define the potential outcomes $Y(M = m)$ as outcome Y for individuals when M is set to *m*. Next, denote the mean of M in group A as m_A and that in group B as m_B. Equalizing M as described, we are now interested in the expected value of Y in group B when M is set to m_A: $E\left(Y^B(M = m_A)\right) - E(Y^A)$. Note that group A does not have potential outcomes notation since the mean of M in group A is already m_A. To reveal the inference problem, we can expand the observed expected value of Y in group B by conditioning on M:

$$E(Y^B) = \sum_M E(Y|M = m, B)P(M = m|B) \quad (2)$$

If we want to estimate $E\left(Y^B(M = m_A)\right)$, we could replace the expected value of Y for those with $M \neq m_A$ with $E(Y|M = m_A)$ and sum over the distribution of the probabilities. Then we would estimate the potential outcome as:

$$E\left(Y^B(M = m_A)\right) = \sum_M E(Y|M = m_A, B)P(M = m|B) \quad (3)$$

However, this quantity is only valid as a counterfactual if the observed outcomes Y for those with $M = m_A$ accurately reflect what the outcomes Y for those with $M \neq m_A$ would be if they were instead set to m_A. This condition is also known as exchangeability (Greenland and Robins 1986). In practice, exchangeability is often a strong assumption, given that there are likely other systematic ways those with different levels of M differ that would affect their value of Y. One solution is to find a set of covariates **C**, that when conditioned on make the assignment of M independent from the potential outcomes. That is, conditional on **C**, the value someone had on M is effectively random.

Now, within the strata defined by **C**, we can use the observed expected value of Y for those with $M = m_A$ as the counterfactual for those with $M \neq m_A$:

$$E(Y(M = m_A)|M \neq m_A, B, C) = E(Y|M = m_A, B, C) \quad (4)$$

Replacing this across the strata of **C** and M, we estimate $E\left(Y^B(M = m_A)\right)$ as:

$$E\left(Y^B(M = m_A)\right) = \sum_M \sum_C E(Y|M = m_A, C, B)P(M = m, C = c|B) \quad (5)$$

While this quantity could be estimated directly there are three specific problems. First, as the number of **C** variables increases, many of the cells will be very sparse, and may not contain a reference individual or individuals for whom the observed Y outcome could be calculated and used

as a counterfactual. The most common way to address this problem is to use individual-level data and specify a parametric model for Y as a function of M and use the model to predict the conditional expectations in equation (5) (Hernan and Robins 2019; Naimi, Cole, and Kennedy 2017; Robins 1986). Second, M might be distributed differently between the two groups, yet this approach only matches the first moment of M across groups, ignoring the role of the distributional differences in M. If our aim is to equalize the entire distribution of M between B and A, we would want to estimate:

$$E\left(Y^B\left(M{\sim}f(M_A)\right)\right) = \sum_M \sum_C E(Y|M{\sim}f(M_A), C, B)P(M{\sim}f(M), C = c|B). \quad (6)$$

Where $f(M_A)$ refers to the distribution of M in population A. This is especially relevant when Y and M have a non-linear relationship, such as when Y is binomially distributed, and Y and M follow a logistic relationship. In such instances, only matching the mean between two populations may not be appropriate for decomposition (we demonstrate this with simple examples in Appendix 2). Finally, the variable type of the mediator may complicate this calculation further. For example, if M is a binomial variable, we face the problem that probabilities are not directly observable for any given individual. To address these latter two issues, we can assign a value for M for each individual in the data based on draws from the distribution of interest. For example, for a continuous M we would match Y in group B to the distribution of Y in group A by drawing continuous values from a normal distribution with a mean and standard deviation given by $f(m_A)$; if M is binomial, we would draw discrete values for M in group B from, π_A, the probability of M in group A. We will move through this process intuitively in two steps.

First, we model Y as a function of M and the **C** variables needed for exchangeability using a parametric model (e.g. if Y is binomial, g( ) could be a logistic function):

$$g\left(E(Y^B|M, C)\right) = \alpha_B + M \cdot \beta_B + \sum C_i \cdot \gamma_B^i \quad (7)$$

Where the conditional expectation function would be:

$$E(Y^B|M,C) = g^{-1}\left(\alpha_B + M \cdot \beta_B + \sum C_i \cdot \gamma^i\right). \quad (8)$$

This model provides a solution to the dimensionality problem discussed above but does not address the issue of how we set the distribution of M to $f(m_A)$ for group B. We match the distribution by drawing a new value of M for each individual in group A directly from $f(m_A)$. For binomial M's, this addresses the issue that probabilities are not directly observable by assigning a discrete value for each individual by drawing from a binomial variable with probability π_A. After matching the distribution across groups, we now estimate the potential outcome using the original model coefficients with the updated counterfactual values of M as:

$$E\left(Y^B\left(M{\sim}f(m_A)\right)\right) = \sum_{C,M} g^{-1}\left(\alpha_B + M^* \cdot \beta_B + \sum C_i \cdot \gamma_B^i\right) P(C = c, M = m|B) \quad (9)$$

where M* indicates that m has been drawn from $f(m_A)$.

At this point, we can form the estimate of $\Delta_{counterfactual}$. Although it may be appealing to directly compare this counterfactual difference to the observed difference, this is potentially problematic because the counterfactual difference is model-based while the observed difference is not. Therefore, differences between the observed and counterfactual contrasts could be due to both the contribution of M and the modeling process. To remove the influence of the model, we first use the model to predict the observed difference using the original data and use this predicted difference as the reference in our estimates of the contribution. This prediction of the observed data is often referred to as the "natural course" estimate. This requires that we predict Y in both groups using the same model, otherwise differences in the model might create differences between the natural course and counterfactual predictions. Depending on the flexibility needed, this could be done by introducing a dummy variable for group A/B into one model, additionally interacting this dummy with M and (some of) C, or separately estimating the model for Y for both groups. For this example,

we will assume that separate models were estimated for both groups. Therefore, the ultimate contrast we now form is based on the following three quantities:

$$E(Y^B): \sum_{C,M} g^{-1}\left(\alpha_B + M \cdot \beta_B + \sum C_i \cdot \gamma_B^i\right) P(C = c, M = m|B) \tag{11}$$

$$E\left(Y^B\left(M{\sim}f(M_A)\right)\right): \sum_{C,M} g^{-1}\left(\alpha_B + M^* \cdot \beta_B + \sum C_i \cdot \gamma_B^i\right) P(C = c, M = m|B) \tag{12}$$

$$E(Y^A): \sum_{C,M} g^{-1}\left(\alpha_A + M^* \cdot \beta_A + \sum C_i \cdot \gamma_A^i\right) P(C = c, M = m|A) \tag{13}$$

To estimate the contribution of M, we also need to specify a contrast for Δ. If, for example, we set Δ to be a simple mean difference then:

$$\Delta Y_{observed} = E(Y^B) - E(Y^A) \tag{14}$$

$$\Delta Y_{counterfactual} = E\left(Y^B\left(M{\sim}f(M_A)\right)\right) - E(Y^A) \tag{15}$$

The contribution of M, as defined above, is then estimated as:

$$\text{Contribution } = 1 - \frac{E\left(Y^B\left(M{\sim}f(M_A)\right)\right) - E(Y^A)}{E(Y^B) - E(Y^A)}. \tag{16}$$

More generally, we can define Δ as any function of Y^A, Y^B, and not just have to rely on collapsible coefficient estimates of contrasts such as the risk difference and risk ratio.

Monte Carlo-Based Estimation

In theory, an analytical solution for the counterfactual decomposition can be found. However, the method outlined above has a few challenges that prevent direct estimation. First, the number of expectations that need to be calculated can increase substantially as the number of confounders and mediators increases. Second, if our goal is to not just to match the first moment of M but rather the entire distribution of M, our method requires identifying and enumerating the different distributions of M present in the data. It is not obvious how to categorize people directly on the basis of their M distribution. Therefore, a more general solution is to simulate entire populations under the observed and counterfactual scenarios and estimate the unconditional expectations directly by taking averages from the simulated data. This is an oft-applied computational procedure for many forms of g-computation (Hernan and Robins 2019; Imai, Keele, and Tingley 2010; Wang and Arah 2015).

Our method can be understood as estimating the contribution of a mediator or several mediators by first generating an entire counterfactual population where the mediator is changed in some way and then comparing moments from this population to a natural-course population that has not been changed. This pseudo-population perspective is powerful because it easily allows for comparisons of not just mean differences but any contrast we can think of (such median differences or a function of population moments such as life expectancy) since we have effectively re-generated an entire micro-population for the observed and counterfactual worlds. Provided that the modeling procedure was flexible enough to allow for subgroup-specific effects, we could also focus on specific subgroups by simply limiting our comparison to specific observations in the pseudo-population. This perspective leads to an algorithmic way of estimating these decompositions that is general enough to be applied to a wide range of scenarios:

Step 1: Estimate relationships in the data by fitting a regression model for the outcome with both the mediator(s) of interest and confounders of the mediator-outcome relationship as covariates.

Step 2: Form the Natural Course Pseudo-Population.

a. Use the estimated model coefficients to simulate the outcome in the different population groups based on the original data. This is the natural course scenario.
b. Using the natural course pseudo-population, form the contrast of interest across groups.

Step 3: Form the Counterfactual Pseudo-Population

a. For the non-reference groups, set the distribution of the mediators by drawing values from the distribution of the mediator in the reference group.
b. Using the original regression coefficients with the counterfactual data, simulate the outcome for each individual.
c. Form the contrast of interest across the counterfactual populations.

Step 4: Estimate the contribution of the mediator(s) to population differences by comparing the contrast of interest in the natural-course and counterfactual pseudo-populations.

Note that in step 1, the regression model can be estimated separately for the various populations being compared or can include dummy variables (potentially interacted with relevant covariates or mediators) for each group. In Step 4 we do not compare the counterfactual and observed contrasts since the difference between those two represents both the difference due to the mediator and the difference due to the model. However, to confirm that the natural course simulations adequately approach the empirical data, we do strongly recommend that the distribution of simulated outcomes of interest for population A is compared to the empirical distribution of the outcome for this population, and likewise for population B. If there is a large discrepancy between the predicted and

observed values, the model's functional form may need to be recalibrated. Simulated values in Steps 2a and 3b should mimic the empirical data; i.e. if the outcome is binomial, we do not predict only probabilities, but assign a 0 or 1 to each individual based on a random draw from a binomial distribution with the predicted probability for that individual. This rule is not needed if the interest is in the (difference in) means of populations. In that case, one can average over values predicted by the underlying regression models, rather than first entering those predicted values as parameters in distributions from which to draw stochastically. For these scenarios, this is a computationally more efficient approach.

Our approach has two types of variability that have to be dealt with separately. First, the assignment of the prediction in step 2b and 3b involves drawing discrete values from a distribution of interest, introducing Monte Carlo error. To reduce this error, either a large pseudo-population could be constructed at the start of Step 2 through data cloning, or Steps 2 and 3 could be performed multiple times, averaging over these steps to produce stables estimate of each contrast, before moving on to Step 4.

Second, because our results are based on a sample, we need to account for sampling variability; this is especially important for the construction of confidence intervals around the estimates. We use a bootstrap procedure to capture this uncertainty, drawing with replacement a fresh sample of size equal to the original data at the start of the steps, conducting the entire analysis k times, and then averaging across the k bootstrap samples to obtain a point estimate and using the 2.5% and 97.5% percentiles for the confidence interval.

Empirical Example 1: Stroke's Contribution to Education Differences in Activities of Daily Living Limitations in Mexico

Data: The Mexican Health and Aging Study

In this section, we demonstrate our approach using data from the 2012 wave of the Mexican Health and Aging Study (MHAS). The MHAS is a nationally representative longitudinal survey of Mexican adults over the age of 50 and their spouses. The MHAS currently has four waves of data (2001, 2003, 2012, and 2015) and contains detailed information on individual and household social and economic characteristics as well as detailed information on individual health including biomarker and anthropometric measurements. For this example, we use data on 13,718 individuals from the 2012 wave of data. Since our focus is on the application of our method, we refer readers to the MHAS cohort profile for further details (Wong, Michaels-Obregon, and Palloni 2015).

Outcome (Y)

Our primary outcome is a binary indicator for whether an individual has any activities of daily living (ADL) disabilities. We classify individuals as ADL disabled if they reported difficulty or inability to do any of the following tasks: eating, bathing, dressing, getting out of a bed, and walking across a room.

Grouping variable and mediator (M)

We classify individuals into three schooling groups: less than primary schooling (LPS), primary schooling (PS), and more than primary schooling (MPS). Our main mediator is stroke, based on evidence that stroke is an important cause of disability (Gresham et al. 1975) and may be unevenly distributed across education groups (Ferri et al. 2011). We measure stroke using a binary indicator for whether an individual ever self-reported having had a stroke.

Confounders (C)

We adjust for a number of important potential confounders of the stroke-disability relationship, including health knowledge and access (measured through the number of times in the previous year an individual saw a doctor or visited a hospital and the size of locality they reside in) and whether an individual ever smoked. Health access in particular is an important confounder since stroke is self-reported and individuals who are more disconnected from health systems may be less likely to be aware of their stroke status.

Contrast

Our contrasts of interest are the differences in the prevalence of ADL disability for the LPS and PS groups relative to the MPS group.

Model

We model the relationship between our main outcome (a binary indicator for ADL disability) and mediator (a binary indicator for stroke) using a logistic regression. Therefore, following the steps we outlined previously, the model we fit for step 1b is specified as follows:

$$logit\big(E(Disability|Stroke, \boldsymbol{C})\big) = \alpha + Stroke \cdot \beta + \boldsymbol{C} \cdot \boldsymbol{\gamma} + \boldsymbol{Schooling} \cdot \boldsymbol{\delta}$$

Where stroke is the binary stroke variable, schooling represents a vector containing two dummies for education (less than primary and primary, making more than primary the reference category), and C represents a vector of covariates described previously. The vector C only contains main terms, but we could easily expand this with interactions if needed.

Note that we assume that the relationship between stroke and disability does not vary across the levels of schooling or the confounders. For this reason, we model differences across groups by simply including dummy variables for the medium and high schooling groups. As an alternative we

could have interacted the schooling dummies with the other variables or estimated separate models for all three schooling groups. For this analysis, we estimate the models and conducted the decomposition separately for men and women. Code for this analysis is available in Supplemental Material [will be available with R package].

Results

Table 1 presents the observed prevalence of ADL disability across schooling groups separately for men and women. For both men and women, there are large differences in ADL disability across schooling groups. For men, 22.6% of individuals with no schooling are ADL disabled compared to only 7.3% of individuals with more than primary schooling. The level of ADL disability is higher for women across all three groups compared to men, with a prevalence of 28.7% among women with no schooling compared 11.4% among women with more than primary schooling. For both sexes, the prevalence of disability for individuals in the some primary schooling group is between the no schooling and more than primary schooling groups.

Table 2 presents the counterfactual-based decomposition results for men (Panel A) and women (Panel B). For men, the natural course estimate of the prevalence of ADL disability is 22.5% for individuals with less than primary schooling compared to just 7.2% for those with more than primary schooling, resulting in a 15.3 percentage point difference. The magnitude of the difference is smaller, but still substantial, between those with primary schooling and those with more than primary schooling (NC diff of 7.7 percentage points). After setting the lower two schooling groups to have the same distribution of stroke as the highest schooling group, these differences shrink to 13.8 and 7.0 percentage points respectively. Applying equation (1), compared to men with more than primary schooling, stroke explains $\left(1 - \frac{0.138}{0.153}\right) \cdot 100\% = 9.7\%$ of the difference in ADL disability

for the no schooling group and $\left(1-\frac{0.070}{0.077}\right)\cdot 100\% = 9.6\%$ for the primary schooling group. Although the magnitude of the natural course difference in ADL disability across schooling groups is larger for women (17 percentage point difference between the no and greater than primary schooling groups and 9 percentage points between the primary and greater than primary schooling groups), the differences do not change as much after equalizing the distribution of stroke. We find that differences in the distribution of stroke only explain around 4% of the difference in ADL disability across schooling groups among women.

Empirical Example 2: Contribution of years of education to childhood socioeconomic differences in fertility

Data: The 1970 British Birth Cohort

In this section, we demonstrate our approach using data from the 1970 British Cohort Study (BCS70) (Elliott and Shepherd 2006; UK Data Archive n.d.). The BCS70 routinely follows around 17,000 individuals born in Great Britain (except Northern Ireland) in a single week in 1970 (UK Data Archive 2016). Beginning in the 26-year follow up, women were asked about their pregnancy history. For this example analysis, we use data on women from the 2008-2009 follow-up wave when the women were 38 years old. We only include those with non-missing baseline and pregnancy follow-up information for a total sample of 3634 women.

Outcome (Y)

We study two closely related outcomes in this study: being childless at age 38 (binomial), and time to first birth (time-to-event). These variables were based on self-reported prior pregnancy histories. For this analysis, we only consider live births.

Grouping variable and mediator (M)

We classify individuals into three childhood socio-economic (SES) groups. To construct the childhood SES groups, we first conduct a principal components analysis on binary indicators for tertiles of parental income when the cohort members were aged 10, tertiles of the respondent's mother's age at delivery, whether their mother was unmarried at the time of delivery, whether the respondent's mother was college educated, whether the respondent's father was college educated, and indicators for occupational classes for both the father and mother. We then create a continuous SES score using the first principal component and classify individuals into tertiles of the score. Our

mediator of interest is the number of years cohort members spent in education, based on theory and evidence that greater time spent in education could both increase the opportunity cost of having a child and the age at which women give birth (Balbo, Billari, and Mills 2013; Becker 1981; Cleland and Wilson 1987).

Confounders (C)

As confounders of the relationship between years spent in education and childbearing, we adjust for region of birth (Scotland, Wales, Northern England, Midlands, Southern England, and London), the age of the mother of the respondent at first childbearing (less than 22, between 22 and 30, and 30 or more), and the number of siblings that the respondent herself had while living in her parental household.

Contrast

Our contrasts of interest are the difference in the prevalence of childlessness relative to the high-income group, and the difference in median age at first birth relative to the high-income group.

Model

We fit separate models for each outcome. We model the relationship between childlessness (a binary indicator) and years spent in education using the following logistic model:

$$logit\big(E(Childlessness|Years\ in\ Edu, \boldsymbol{C})\big) = \alpha + Years\ in\ Edu \cdot \beta + \boldsymbol{C} \cdot \boldsymbol{\gamma} + \boldsymbol{SES} \cdot \boldsymbol{\delta}$$

Where Years in Edu is the count of number of years in education, SES represents a vector containing two dummies (second and third SES tertiles), and C represents the vector of confounders described previously. We model the relationship between age at first birth and years in education using a Poisson regression with the same variables on the right-hand side:

$$log\big(E(Age\ at\ First\ Birth|Years\ in\ Edu, \boldsymbol{C})\big) = \alpha + Years\ in\ Edu \cdot \beta + \boldsymbol{C} \cdot \boldsymbol{\gamma} + \boldsymbol{SES} \cdot \boldsymbol{\delta}$$

For this outcome, we only examine age at first birth among women with at least one child. Similar to the assumption made in Example 1, we assume that the relationship between years in education and the childlessness (and age at first birth) does not vary across levels of schooling or the confounders. Code for this analysis is available in Supplemental Material [will be available with R package].

Results

Table 3 presents the observed percentage of women who are childless at age 38 and the median age at first birth among those that have had at least one birth across childhood SES groups. Women with higher childhood SES are more likely to be childless and have an older age at first birth. For childlessness, 31.4% of women with high childhood SES are childless at age 38 compared to only 22.9% among women with low childhood SES. Among women who have had at least one birth, women with high childhood SES have a median age at first birth that was three years greater than for women with low childhood SES (28.0 vs. 25.0). For both outcomes, women from the middle childhood SES group are between the low and high groups.

Table 4 presents the counterfactual-based decomposition results for childlessness (Panel A) and age at first birth (Panel B). There is an 8.5 percentage point difference in the natural course difference in percentage childless at age 38 between women with high and low childhood SES, and a 4.2 percentage point difference between women with high and medium childhood SES. After setting the lower two childhood SES groups to have the same distribution of total years spent in education as the high childhood SES group, these differences reduce substantially (5.8 percentage points and 2.1 percentage points for the low and medium childhood schooling groups respectively). Based on

this change, we estimate that 32.2% of the difference in childlessness between high and low childhood SES groups and 50.2% of the difference in childlessness between the high and medium childhood SES groups is due to differences in total years spent in education.

We find similarly high levels of mediation for median age at first birth. For example, there is a 3.2-year natural course difference in median age at first birth between the high and low childhood SES groups. After setting the low childhood SES group to have the same years spent in education as the high SES group, this difference reduces to 2.6 years. Therefore, we estimate that differences in total years spent in education between the two groups is responsible for 18.7% of the 3.2-year difference in median age at first birth between those with high and low childhood SES.

Discussion

We introduce a generalized yet easily applied procedure for implementing Jackson and VanderWeele's (2018) (henceforth JVW) counterfactual decomposition for decomposing social or population differences in a wide range of outcomes (Jackson and VanderWeele 2018). This method allows for the decomposition of group differences into the contribution of mediating or explanatory variables within a pre-specific causal structure. Our approach has two important advantages that improve its applicability to questions common in demography and the populations sciences. First, our method provides an implementational advance since it does not rely on closed form approximate solutions and can be applied to any parametric model without having to derive any decomposition equations. Second, our method provides a theoretical contribution to JVW's counterfactual decomposition by allowing for the decomposition of population differences in any function of the outcome at any point of the outcome distribution, not just risk ratios or mean differences. This is especially relevant for the population sciences, where comparisons between populations are often done using non-linear or non-mean-based indicators, such as differences in life expectancy, age-standardized rates, and median time-to-events.

Other decompositions

Our method is an operationalization of the JVW decomposition, but it is conceptually related to several other existing decompositions from the social sciences. The intuition behind our approach is similar to many decompositions in demography that evaluate the contribution of a mediator by swapping the distribution of the mediator in one population with that of a reference and then calculating the change in the contrast between the observed and re-weighted populations (Andreev et al. 2002; Chevan and Sutherland 2009; Geruso 2012; Gupta 1978, 1993; Kitagawa 1955). In general, however, these demographic decompositions have two principle limitations that our

method overcomes. The first is that they are often purely non-parametric, requiring that mediator variables be split into discrete bins – a choice that is often arbitrary and can lead to issues of sparseness. This problem is magnified when considering multiple mediators since the methods now require splitting the joint density of the mediators into bins (Gupta 1993). The second, related, limitation is that these decompositions can only adjust for confounders through stratification. This requires that the confounder be split into bins, the decomposition conducted within each stratum of the confounder (complete with the joint density also split into bins), then the strata-specific results be re-aggregated across the bins of the confounder. Both these two issues rapidly lead to dimensionality problems that are solved in our approach through the use of parametric models. One important difference between our approach and traditional demographic decompositions, however, is that our approach necessitates micro-level data and often will not be possible to conduct with aggregate data. Therefore, this decomposition is not intended to replace traditional decompositions which are often applied to aggregate data, but rather complement these decompositions with more concrete counterfactual interpretations when micro-level data are available.

Our approach is also similar to the Oaxaca-Blinder and related decompositions (Blinder 1973; Oaxaca 1973). JVW mathematically prove that their decomposition is equal to the Oaxaca-Blinder decomposition when the Oaxaca-Blinder decomposition includes the same set of confounders, none of the confounders are also on the causal pathway between the mediator and the outcome, and when the outcome can be modeled using an OLS regression. We refer readers to the original paper for these details but present additional distinctions that arise from our implementation. First, the Oaxaca-Blinder was created to decompose differences in a continuous variable. While non-linear implementations exist, they rely on additional assumptions and require the analyst to derive a separate set of equations for each distribution of the outcome variable (Bauer and Sinning 2008). Second, Oaxaca-Blinder decompositions are only applicable for decomposing mean

differences between two populations, ignoring other population contrasts. Finally, Oaxaca-Blinder decompositions only consider mean differences in the mediator between the two groups and ignore the role of the entire mediator distribution. As mentioned previously, this can create incorrect contribution estimates, especially when working with non-linear outcomes. By working with simulations and the Monte Carlo approach, our method does not have these three major limitations.

A common but often incorrectly used method in the social sciences is to estimate a regression including the group as a categorical variable and then observing how the coefficient on the group variable changes when potential mediators are added. This method, however, is only valid for outcome/contrast combinations that are collapsible (e.g. mean differences and risk ratios) and is therefore limited in applicability (Greenland, Robins, and Pearl 1999).

Limitations and other considerations

We use Monte Carlo methods to provide a general solution to calculating expectations and bootstrap to estimate standard errors. Both these procedures are computationally costly and require high processing power and/or long computational times. Furthermore, since we match mediator distributions non-parametrically (by drawing from the reference group's mediator distribution) we recommend a large sample size to ensure that the sampled distribution approximates the real population's distribution. This compounds the computational cost issue, as larger data sizes increase computational requirements. However, our approach can still be used if only a small sample size is available by assuming a parametric distribution for the mediator. This approach, however, will result in bias if the assumed distribution is incorrect.

Results from our approach are only valid if the underlying assumption of no unmeasured confounding is correct. Since the counterfactual decomposition approach is meant to answer the question "to what extent do differences in the distribution of M between groups of X account for

the association between X and Y", we are primarily concerned about confounders of the mediator-outcome (M-Y) pathway and not the group-outcome (X-Y) pathway (the causal effect of the group variable on the outcome). If the interest is in decomposing the causal effect of X on Y, then confounders of the X-Y and X-M pathways should also be adjusted for.

In addition to including the correct confounders, we also have to correctly estimate the functional form of the mediator and confounders. This is because our approach relies on using the estimated models to predict values for individuals; if the functional forms are incorrect, there is a chance that the predictions will not accurately represent the observed data even if all the correct confounders are included. An advantage of our approach is that it allows for any functional form for the underlying models. We recommend that natural course predictions are compared with corresponding values from the empirical data to ensure that there are not large prediction errors. Although this does not prove that the model is correctly specified, it helps avoid grossly misspecified models.

One conceptual issue that may arise is a lack of common support (also known as positivity) of the mediator distribution across two groups. For example, suppose we are interested in equalizing the distribution of education between women with high and low childhood SES. If the low SES group has total education values of 6 to 9 and the high SES group has values ranging from 6 to 12, it cannot be determined from the data how the low education group would respond to having education values above 9; in such a case, when setting the education distribution of the low SES group to that of the high SES group, one may be forced to assume that the relationship between total education values above 9 in the low SES group is the same as that of the high SES group.

Finally, the selectivity of the sample must be considered in the interpretation of the results. In Example 1, we estimate the contribution of stroke to schooling differences in disability. To do so, we take a sample of individuals aged 50+ and estimate from them the relationship between stroke

and disability. When we assign the stroke distribution of the highly educated to the low educated, we are not answering the question 'what if low educated individuals in general would have had the stroke distribution of the highly educated', but instead 'what if the low educated individuals who survived to age 50 would have had the stroke distribution of the highly educated'. This distinction is important because some individuals in the population of interest who were low educated may have died of stroke or other causes before age 50 and would have survived (and thereby potentially entered the sample) if the low educated group had the stroke distribution of the highly educated at all ages. This issue is especially problematic if the low educated who died of stroke before age 50 had a different relationship between stroke and disability than the survivors, which is also known as survivorship bias or selection bias. If the interest is in the life course contribution of stroke, a sample covering all age groups should be taken.

As with most decompositions, the estimate of the contribution is dependent on both the contrast of interest and the reference group used for the counterfactual. For example, looking at the absolute difference in means between two groups and how that difference would change if mediator values were equalized will result in a different estimate of the contribution than if the relative difference between the two groups had been used as a contrast instead. Similarly, setting the stroke distribution of the highly educated to the stroke distribution of the low educated would result in a different estimate of the contribution than setting the stroke distribution of the low educated to that of the high. We do not view this as a true limitation, as it gives the decomposition a clear counterfactual interpretation and forces the analyst to substantively motivate these choices. For example, in our stroke analysis we chose to look at the difference in means between schooling groups because we find absolute differences in disability prevalence more interpretable than relative ones. Furthermore, we set the stroke distribution of the no schooling group (where stroke prevalence is high) to that of the highly educated group (where stroke prevalence is low) since a

realistic intervention effort to reduce inequality would work by trying to reduce the stroke prevalence of the no schooling group rather than increase stroke in the higher schooling groups.

Conclusions

Decomposing the sources of differences in health and other outcomes is a key research endeavor in demography and other population sciences. We introduce a flexible implementation of counterfactual decomposition that builds on and generalizes the rich existing body of work on decomposition methods in the health and social sciences. Our approach is highly flexible and easily implemented way of estimating decompositions that are grounded in potential outcomes and counterfactual theory.

Appendix 1: Decomposition results when the mediator – outcome relationship is confounded

In this appendix, we show an example of how traditional demographic decompositions will provide incorrect contribution estimates when the relationship between the mediator and outcome of interest is confounded. For this example, we demonstrate this using a simple Kitagawa decomposition, but the same finding would hold for other decompositions that do not explicitly account for confounding.

Hypothetical Question

What is the contribution of systolic blood pressure (BP) to the difference in disability between two groups (group 1 and 2)?

Confounding

For this example, we will assume that systolic blood pressure and disability share the common cause of waist circumference (the confounder). This is not an unrealistic assumption, as many studies have shown that waist circumference affects BP and affects disability through other causes such as arthritis and diabetes.

Generating the data

We begin by drawing two separate distributions of waist circumference for groups 1 and 2 with a mean difference of 10 cm between groups:

```
# group size
g.size <- 1000000

# Waist circumference
waist.g1 <- rnorm(g.size, mean = 110, sd = 5)
waist.g2 <- rnorm(g.size, mean = 100, sd = 4)
```

We then create blood pressure values for each group using the following expressions:

```
# BP as a function of waist (using 0.35 as the relationship between
waist and BP)
bp.g1 <- 130 + 0.35*waist.g1 + rnorm(g.size, mean = 0, sd = 5)
bp.g2 <- 120 + 0.35*waist.g2 + rnorm(g.size, mean = 0, sd = 6)
```

Finally we express the probability of disability as a logistic function of BP and waist circumference, and draw specific values of disability for individuals in both groups from these disability probabilities:

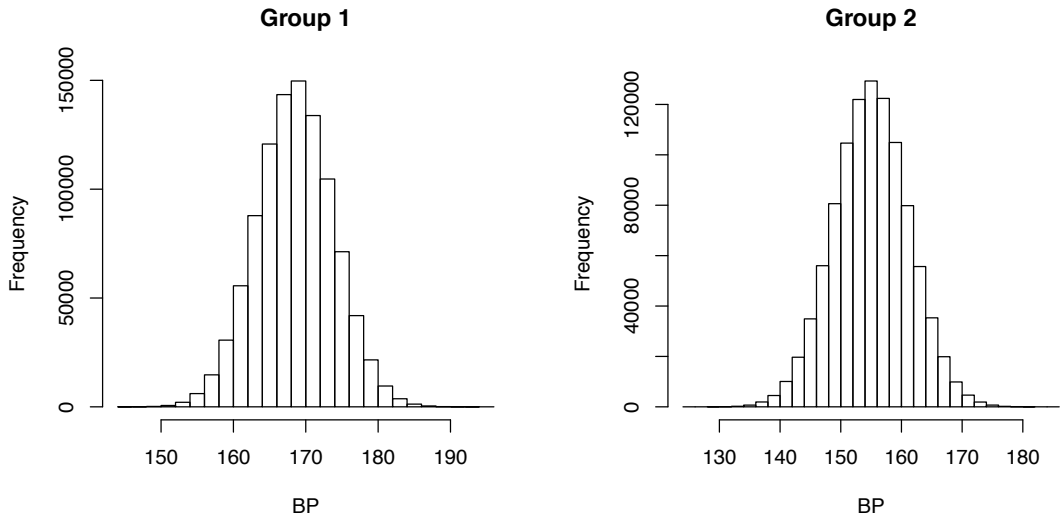
```
#Turn into data
toy <- data.frame(rbind(cbind(waist.g1, bp.g1,
rep(1,g.size)),cbind(waist.g2, bp.g2, rep(2,g.size))))
colnames(toy) <- c("waist","bp","group")

# Disability as a function of both BP and waist
```

```
expit <- function(x) exp(x)/(1+exp(x))
toy$prob <- expit(-3+0.00499*toy$bp+0.00995*toy$waist)
toy$disability <- rbinom(2*g.size, size = 1, prob = toy$prob)
```

Generated Data

This results in the following data for BP:



With a prevalence of disability of 25.6% in group 1 and 22.6% in group 2.

Conducting the Decompositions

We first conduct a Kitagawa decomposition of disability between groups 1 and 2 by breaking up BP into bins of size 5 between 120 and 190 mmHg systolic BP. We then conduct two versions of the counterfactual decomposition, one where we did not control for confounding from waist circumference and one where we did. Since our decomposition requires specifying a counterfactual question, we estimate the answer to the question "How much smaller would the difference in the prevalence of disability between groups 1 and 2 be if they both had the same distribution?" To be consistent with the Kitagawa, we assign that distribution to be the average of the two distributions.

Results

Decomposition Table

	Kitagawa*	CFL Decomp no confounders	CFL Decomp with confounders
Contribution of BP	58%	58%	40%

*We show the contribution of difference in the distribution of BP term from the Kitagawa decomposition.

The results reveal that if we did not control for confounding for waist circumference, we would drastically overestimate the contribution of differences in the distribution of BP to the difference in disability between groups. This occurs because part of the contribution is driven by differences in

the distribution of waist circumference and not BP. Secondly the results show that the in absence of controls for confounding the CFL decomp is equivalent to the Kitagawa, which is what we would expect based on results from JVW.

Could we have controlled for waist circumference in the Kitagawa Decomposition?

Technically there is nothing preventing us from controlling for waist circumference in the Kitagawa decomposition. However, doing so would require splitting waist circumference into bins, separately conducting the Kitagawa decomposition within the bins of waist circumference, and then reaggregating the results across bins. Even with just one confounder this rapidly creates large dimensionality issues, a problem that becomes even worse with multiple confounders.

Full R Code

Attached as appendix1code.R

Appendix 2: Decomposition results when the distribution of the mediator is not properly matched

In this example, we demonstrate that when there are non-linear relationships between the mediator and the outcome, simply matching the mean across populations will produce incorrect estimates of decomposition.

We demonstrate this with hypothetical data with a continuous mediator (M), a binomial outcome (Y) that is a function of the continuous mediator, and two different groups. We consider three scenarios:

1. Both groups have the same distribution of the mediator but with a different mean. This would correspond to a simple mean shift in the distribution across groups.

2. The groups have the same mean of the mediator but the distribution around this mean is different across the groups.

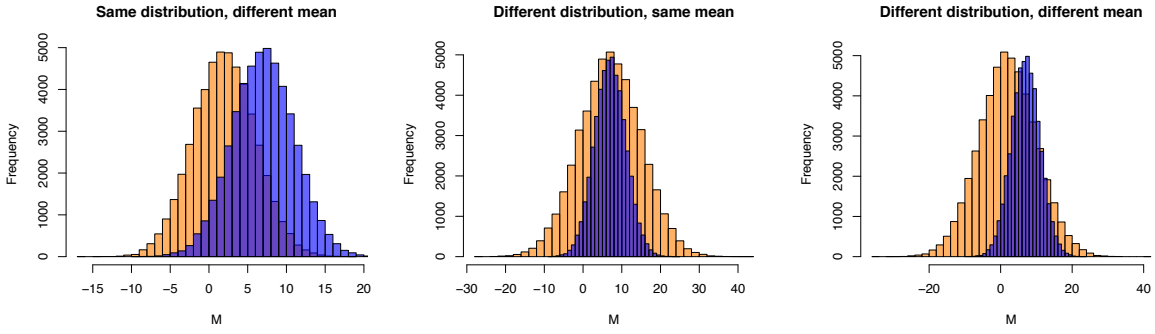
3. The groups have a different mean and a different distribution of the mediator.

For each of these scenarios we calculate our counterfactual decomposition two separate ways. We first correctly match the distributions and second, we just match the mean. The counterfactual question we examine is, "How much would the difference in the outcome Y between groups 1 and 2 decrease if group 1 had the same distribution of the mediator M as group 2?"

The code to generate the data and produce these results is given at the end of this appendix.

Results

Here are the three scenarios we consider:



For these scenarios, the decomposition results are:

Decomposition Results – Percent contribution of M to differences in Y between groups 1 and 2

	Matched distributions	Matched means
Same distribution, different mean	87%	110%

Different distribution, same mean	43%	140%
Different distribution, different mean	84%	115%

Even when the distributions are truly just separated by a mean shift (the first scenario), the non-linear relationship between M and Y results creates a situation where just matching the mean of M and not the entire distribution, substantially over-estimates the contribution of M. This discrepancy can also be seen when the two populations have the exact same mean but a different distribution.

R Code

Attached as appendix2code.R

References

Andreev, Evgueni M., Vladimir M. Shkolnikov, and Alexander Z. Begun. 2002. "Algorithm for Decomposition of Differences between Aggregate Demographic Measures and Its Application to Life Expectancies, Healthy Life Expectancies, Parity-Progression Ratios and Total Fertility Rates." *Demographic Research* 7:499–522.

Arriaga, Eduardo E. 1984. "Measuring and Explaining the Change in Life Expectancies." *Demography* 21(1):83–96.

Balbo, Nicoletta, Francesco C. Billari, and Melinda Mills. 2013. "Fertility in Advanced Societies: A Review of Research." *European Journal of Population/Revue Européenne de Démographie* 29(1):1–38.

Bauer, Thomas K. and Mathias Sinning. 2008. "An Extension of the Blinder–Oaxaca Decomposition to Nonlinear Models." *AStA Advances in Statistical Analysis* 92(2):197–206.

Becker, Gary Stanley. 1981. *A Treatise on the Family*. Harvard university press.

Blinder, Alan S. 1973. "Wage Discrimination: Reduced Form and Structural Estimates." *Journal of Human Resources* 436–455.

Chevan, Albert and Michael Sutherland. 2009. "Revisiting Das Gupta: Refinement and Extension of Standardization and Decomposition." *Demography* 46(3):429–449.

Cleland, John and Christopher Wilson. 1987. "Demand Theories of the Fertility Transition: An Iconoclastic View." *Population Studies* 41(1):5–30.

Elliott, Jane and Peter Shepherd. 2006. "Cohort Profile: 1970 British Birth Cohort (BCS70)." *International Journal of Epidemiology* 35(4):836–843.

Ferri, Cleusa P., Claudia Schoenborn, Lalit Kalra, Daisy Acosta, Mariella Guerra, Yueqin Huang, KS Jacob, Juan J. Llibre Rodriguez, Aquiles Salas, Ana Luisa Sosa, and others. 2011. "Prevalence of Stroke and Related Burden among Older People Living in Latin America, India and China." *Journal of Neurology, Neurosurgery & Psychiatry* 82(10):1074–1082.

Geruso, Michael. 2012. "Black-White Disparities in Life Expectancy: How Much Can the Standard SES Variables Explain?" *Demography* 49(2):553–574.

Greenland, Sander and James M. Robins. 1986. "Identifiability, Exchangeability, and Epidemiological Confounding." *International Journal of Epidemiology* 15(3):413–419.

Greenland, Sander, James M. Robins, and Judea Pearl. 1999. "Confounding and Collapsibility in Causal Inference." *Statistical Science* 29–46.

Gresham, Glen E., Therese E. Fitzpatrick, Philip A. Wolf, Patricia M. McNamara, William B. Kannel, and Thomas R. Dawber. 1975. "Residual Disability in Survivors of Stroke—the Framingham Study." *New England Journal of Medicine* 293(19):954–956.

Gupta, Prithwis Das. 1978. "A General Method of Decomposing a Difference between Two Rates into Several Components." *Demography* 15(1):99–112.

Gupta, Prithwis Das. 1993. *Standardization and Decomposition of Rates: A User's Manual*. US Department of Commerce, Economics and Statistics Administration, Bureau ….

Hernan, Miguel A. and James M. Robins. 2019. *Causal Inference*. CRC Boca Raton, FL:

Ho, Jessica Y. 2013. "Mortality under Age 50 Accounts for Much of the Fact That US Life Expectancy Lags That of Other High-Income Countries." *Health Affairs* 32(3):459–467.

Horiuchi, Shiro, John R. Wilmoth, and Scott D. Pletcher. 2008. "A Decomposition Method Based on a Model of Continuous Change." *Demography* 45(4):785–801.

Imai, Kosuke, Luke Keele, and Dustin Tingley. 2010. "A General Approach to Causal Mediation Analysis." *Psychological Methods* 15(4):309–34.

Jackson, John W. and Tyler J. VanderWeele. 2018. "Decomposition Analysis to Identify Intervention Targets for Reducing Disparities." *Epidemiology* 29(6):825–835.

Kitagawa, Evelyn M. 1955. "Components of a Difference between Two Rates." *Journal of the American Statistical Association* 50(272):1168–1194.

Kittner, Steven J., Lon R. White, Katalin G. Losonczy, Philip A. Wolf, and J. Richard Hebel. 1990. "Black-White Differences in Stroke Incidence in a National Sample: The Contribution of Hypertension and Diabetes Mellitus." *Jama* 264(10):1267–1270.

Martikainen, Pekka, Pia Mäkelä, Riina Peltonen, and Mikko Myrskylä. 2014. "Income Differences in Life Expectancy: The Changing Contribution of Harmful Consumption of Alcohol and Smoking." *Epidemiology* 25(2):182–90.

Naimi, Ashley I., Stephen R. Cole, and Edward H. Kennedy. 2017. "An Introduction to g Methods." *International Journal of Epidemiology* 46(2):756–62.

Nandi, Arijit, M. Maria Glymour, and SV Subramanian. 2014. "Association among Socioeconomic Status, Health Behaviors, and All-Cause Mortality in the United States." *Epidemiology* 25(2):170–177.

Oaxaca, Ronald. 1973. "Male-Female Wage Differentials in Urban Labor Markets." *International Economic Review* 693–709.

Robins, James. 1986. "A New Approach to Causal Inference in Mortality Studies with a Sustained Exposure Period—Application to Control of the Healthy Worker Survivor Effect." *Mathematical Modelling* 7(9–12):1393–1512.

Steckel, Richard H. and Garrett Senney. 2015. *Historical Origins of a Major Killer: Cardiovascular Disease in the American South. Working Paper*. 21809. National Bureau of Economic Research.

UK Data Archive. n.d. *1970 British Cohort Study*. University of Essex, Instiute for Social and Economic Research.

Wang, Aolin and Onyebuchi A. Arah. 2015. "G-Computation Demonstration in Causal Mediation Analysis." *European Journal of Epidemiology* 30(10):1119–1127.

Wong, Rebeca, Alejandra Michaels-Obregon, and Alberto Palloni. 2015. "Cohort Profile: The Mexican Health and Aging Study (MHAS)." *International Journal of Epidemiology* 46(2):e2–e2.

Table 1 Observed crude prevalence of activities of daily living limitations disability by schooling groups, adults ages 50+, Mexican Health and Aging Study, 2012

	No schooling	Some primary schooling	> Primary Schooling
Men	0.226	0.150	0.073
Women	0.287	0.206	0.114

Notes: Activities of daily living limitation disability was measured if an individual reported difficulty or inability for any of the following tasks: eating, bathing, dressing, getting out of a bed, and walking across a room. The some primary schooling group also includes individuals who completed primary schooling but received no further schooling.

Table 2 Estimates of the contribution of stroke to schooling differences in activities of daily living limitations in Mexico using the counterfactual decomposition method, Mexican Health and Aging Study, 2012.

Panel A: Men

	Natural Course Prevalence	Counterfactual Prevalence	Natural Course Difference	Counterfactual Difference	Percent Mediated
> Primary	0.072	0.072			
	(0.061,0.084)	(0.061,0.084)			
Some primary	0.150	0.142	0.077	0.070	9.6%
	(0.137,0.161)	(0.131,0.153)	(0.060,0.093)	(0.053,0.084)	(5.9%,14.4%)
No schooling	0.225	0.210	0.153	0.138	9.7%
	(0.196,0.253)	(0.182,0.237)	(0.123,0.184)	(0.109,0.168)	(6.1%,14.0%)

Panel B: Women

	Natural Course Prevalence	Counterfactual Prevalence	Natural Course Difference	Counterfactual Difference	Percent Mediated
> Primary	0.113	0.113			
	(0.101,0.130)	(0.101,0.130)			
Some primary	0.206	0.201	0.092	0.088	4.7%
	(0.195,0.218)	(0.191,0.214)	(0.074,0.109)	(0.070,0.106)	(2.8%,7.5%)
No schooling	0.286	0.279	0.173	0.166	4.3%
	(0.265,0.310)	(0.257,0.303)	(0.147,0.202)	(0.139,0.193)	(2.5%,6.4%)

Table 3 Observed proportion of women who are childless at age 38 and median age first birth among women who have had at least one birth across childhood SES groups, 1970 British Cohort Study 38-year follow-up.

	Low Childhood SES	Middle Childhood SES	High Childhood SES
Childlessness	0.229	0.271	0.314
Age at first birth	25.0	26.0	28.0

Notes: Childhood SES groups are based on tertiles of a continuous core based on several parental characteristics and parental income when the women were aged 10.

Table 4 Estimates of the contribution of total years of schooling to childhood socioeconomic status differences in every having a birth and median age at first birth among those who have had a birth using the counterfactual decomposition method, 1970 British Cohort Study, 1970-2008

Panel A: Any birth

	Natural Course Percentage	Counterfactual Percentage	Natural Course Difference	Counterfactual Difference	Percent Mediated
High Childhood SES	0.314	0.314			
	(0.290,0.341)	(0.290,0.341)			
Middle Childhood SES	0.271	0.293	0.042	0.021	50.2%
	(0.246,0.299)	(0.266,0.322)	(0.001,0.078)	(-0.019,0.057)	(14.1%,297.3%)
Low Childhood SES	0.229	0.256	0.085	0.058	32.2%
	(0.204,0.252)	(0.229,0.282)	(0.052,0.120)	(0.021,0.095)	(17.4%,61.7%)

Panel B: Age at first birth among those who had a birth

	Natural Course Median Age	Counterfactual Median Age	Natural Course Difference	Counterfactual Difference	Percent Mediated
High Childhood SES	28.0	28.0			
	(27.9,28.3)	(27.9,28.2)			
Middle Childhood SES	26.0	26.7	-2.0	-1.3	34.8%
	(25.9,26.2)	(26.1,27.0)	(-2.3,-1.8)	(-1.9,-1.0)	(7.4%,50.0%)
Low Childhood SES	24.9	25.4	-3.2	-2.6	18.7%
	(24.2,25.0)	(25.0,26.0)	(-3.9,-3.0)	(-3.1,-2.0)	(5.1%,33.1%)