Educational note: Causal decomposition of population health differences using Monte Carlo integration and the g-formula

Nikkil Sudharsanan
Heidelberg Institute of Global Health
Heidelberg University
*Corresponding author

Maarten J. Bijlsma
Laboratory of Population Health
Max Planck Institute for Demographic Research

Summary

One key objective of the population health sciences is to understand why one social group has different levels of health and well-being compared to another. While several methods have been developed in economics, sociology, demography, and epidemiology to answer these types of questions, a recent method introduced by Jackson and VanderWeele (2018) provided an update to decompositions by anchoring them within causal inference theory. In this paper, we demonstrate how to implement the causal decomposition using Monte Carlo integration and the parametric g-formula. Causal decomposition can help to identify the sources of differences across populations and provide researchers a way to move beyond estimating inequalities to explaining them and determining what can be done to reduce health disparities. Our implementation approach can easily and flexibly be applied for different types of outcome and explanatory variables without having to derive decomposition equations and can also decompose functions of outcomes, such as period life expectancy, that are not based around a simple comparison of means or proportions. We describe the concepts of the approach and the practical steps and considerations needed to implement it. We then walk through a worked example where we investigate the contribution of smoking to sex differences in mortality in South Korea using two different outcomes and contrasts: the age-adjusted mortality risk ratio and the absolute difference in period life expectancy. For both examples, we provide both pseudocode and R code using our package, *cfdecomp*. Ultimately, we outline how to implement a very general decomposition algorithm that is grounded in counterfactual theory but still easy to apply to a wide range of situations.

Keywords: decomposition; causal inference; Monte Carlo; parametric g-formula; population models; health disparities.

Key messages

- Causal or counterfactual-based decomposition methods are of growing importance in epidemiology and the population health sciences.
- We develop and demonstrate a highly flexible implementation of the causal decomposition that is grounded in counterfactual theory but still easy to apply to a wide range of questions without having to derive specialized decomposition equations.
- We demonstrate how to use our decomposition algorithm to estimate the contribution of smoking to sex differences in two different summary mortality outcomes in South Korea, finding that smoking explains 27% of the male mortality advantage at ages 50 and above.

Introduction

A central aim of the population health sciences is to understand why one social group has different levels of health and well-being compared to another. Recent examples of this question include understanding why Hispanics have worse congenital heart disease outcomes compared to non-Hispanics (1), why adult mortality is higher in urban compared to rural Indonesia (2), and why poorer individuals in Finland have higher mortality compared to more affluent individuals (3). By identifying the sources of differences across populations, these studies provide an important first step for determining what can be done to reduce health disparities.

Decomposition analyses are one of the key tools for understanding the sources of differences in an outcome between groups and can help to move researchers from estimating to explaining health inequalities. At their core, decomposition analyses seek to determine how much of an observed difference in an outcome between two groups is due to the differing distribution of specific causes of that outcome between the groups. For example, in the example above on Finland, researchers may ask "how much of the mortality difference between rich and poor individuals is due to the higher prevalence of smoking among poorer compared to richer individuals?"

While such questions may sound like mediation analysis (4–8), there is a key difference between mediation and decomposition. In a causal mediation analysis, we would first estimate the causal effect of poverty on mortality, and then identify how much of this effect is driven through poverty's causal effect on smoking. In decomposition analysis, on the other hand, we are interested in how much smoking contributes to observed differences in mortality between poor and non-poor and are agnostic to how much of the difference in smoking between poor and non-poor is due to the causal effect of smoking and how much due to confounding causes. This crucial difference (depicted graphically using DAGs in **Figure 1**) has consequences for the analytical approach to be taken and requires fewer confounding variables to be accounted for. Importantly, in a decomposition analysis,

since we are not attempting to estimate the causal effect of the group variable (the exposure in a mediation in analysis), we do not have to contend with the open issue of whether causal effects can be estimated for non-manipulable characteristics such as race (9).

Various methods have been developed across disciplines for conducting decomposition analyses. Regression decompositions, such as the Oaxaca-Blinder (OB) decomposition (10,11) and its nonlinear extensions (12,13), use individual-level data and are employed frequently in economics and sociology (14), while approaches using aggregate level data are common in demography (15–18). Recent advances in epidemiology provide a new perspective to decompositions, situating them in causal inference and counterfactual theory (2,9,19,20). Among these, Jackson and VanderWeele (2018), provide an important advance by framing decomposition analyses around interventions to reduce disparities, where the importance of specific characteristics to differences between populations is evaluated through hypothetical intervention scenarios to equalize these characteristics between groups (19).

In this paper, we demonstrate a simple way to implement the counterfactual decomposition using parametric models and Monte Carlo integration. We focus on a worked example that asks, "How much of the observed sex-difference in mortality in South Korea is due to the higher prevalence of smoking among men compared to women?" and demonstrate how to decompose sex-differences in two different summary contrasts: the age-adjusted 1-year mortality risk ratio between men and women and the absolute difference in period life expectancy at age 50 between men and women. Our approach can be easily applied within common statistical packages or implemented with our R package, *cfdecomp* (21). Our approach is based on a straightforward algorithm for estimating counterfactual decompositions for any type of outcome distribution without having to derive decomposition equations. As we will demonstrate, the use of Monte Carlo integration has the added advantage of allowing us to decompose complicated summary measures such period life expectancy, which because

they are not a simple contrast of population means or proportions, are more challenging to decompose with closed-form approaches.

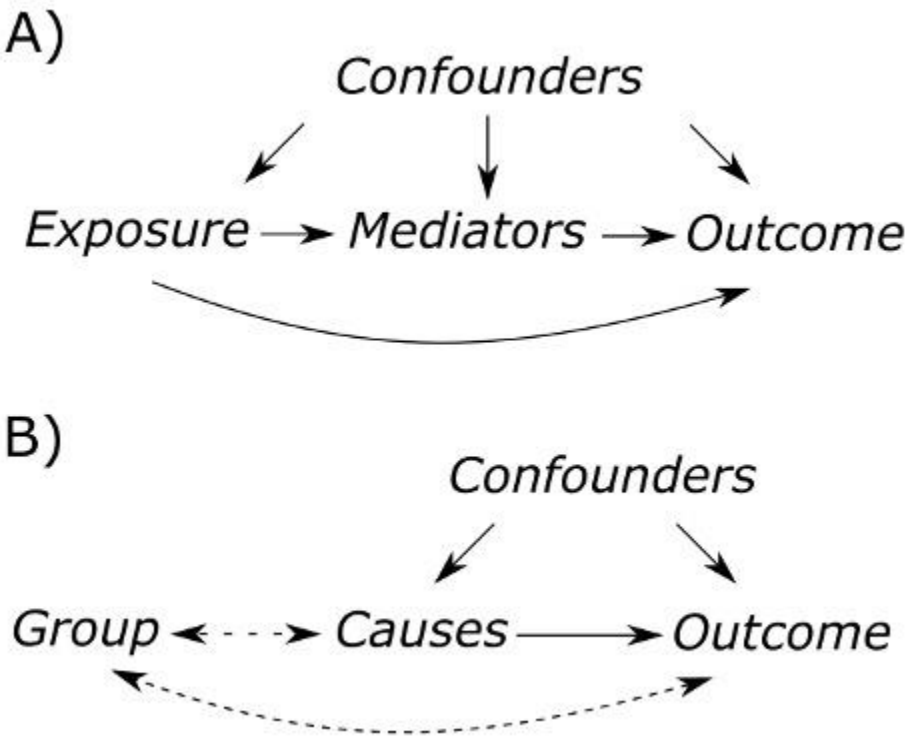

Figure 1. Directed Acyclic Graphs (DAGs) of typical Causal Mediation and Decomposition compared.

A Counterfactual Approach to Decomposition

Concepts

We motivate and develop our approach through the question, "what is the contribution of smoking to sex differences in mortality in South Korea?" We adopt a counterfactual perspective and define "contribution" by asking, "how large would the difference in mortality be if men and women had an equal smoking prevalence?"

Our first main step is to specify exactly what level of smoking prevalence we are equalizing men and women to. When the relationship between an outcome (such as mortality) and a mediator (such as smoking) is non-linear, this choice can affect the contribution estimate (18). Therefore, the choice of the reference distribution should be informed by substantive concerns (e.g. what makes sense from a policy perspective?) and inferential concerns (e.g. certain values may be outside the range observed in the data and should therefore be avoided). We choose to set men to have the smoking prevalence of women, since this maps to a clear intervention that public health policy makers may seek to achieve.

The second main step is to specify a summary population measure. This is the measure that we will use to compare the mortality of men and women in South Korea. We consider two related summary population measures: the age-adjusted one-year risk of death and period life expectancy at age 50 (period life expectancy is a function of the 1-year risks of death across ages).

Third, we need to specify contrasts of these summary measure between men and women (i.e. how are we going to compare the two summary measures?). For the one-year risk of death, we consider the risk ratio for men relative to women (adjusted for age) and for period life expectancy, we consider the absolute difference between men and women.

Based on steps 1 to 3, we can construct our estimate of the "contribution" of smoking by seeing how much the difference in the summary measures between men and women reduces when

we set men to have the same smoking prevalence as women. For example, for life expectancy, we would compare the absolute difference in life expectancy between men and women in the observed data to the difference in life expectancy between men and women in a counterfactual world where we set men to have the same smoking levels as women. We could then estimate the contribution of smoking as the percent reduction in the male-female life expectancy disparity Note that this contribution is not bounded between 0 and 1 and could result in negative contributions or contributions greater than 100%. This is not an issue however; this situation occurs in both mediation and decomposition analyses when the indirect effect (the association via the mediators) or the direct effect (the association not via the mediators) are of opposite signs, and hence partially cancel each other out in the total effect. Indeed, many recent papers using mediation and decomposition analyses have found contribution estimates above 100 or below 0 (2,19,22). Contribution estimates below 0 or above 1 could also occur due to imprecision in the underlying estimates. For this reason, it is important to present and interpret such estimates with their accompanying standard error. In **Appendix 3**, we provide a more general formal exposition of the causal decomposition.

Parametric Modeling and Monte Carlo-Based Estimation

The core estimand in our decomposition is the counterfactual summary measure of mortality for men if they were set to have the same smoking distribution as women. Estimating this counterfactual requires (1) a way to match the smoking distribution between men and women, and (2) a way to re-estimate mortality as a function of the new smoking distribution. Importantly, since we are interested in the effect of changing the level of smoking on mortality, our approach to re-estimating mortality needs to adjust for the confounders of the smoking-mortality relationship. Our solution to these two issues is to use the parametric g-formula and Monte Carlo integration (7,23–25). This entire approach can be estimated by following a straightforward algorithm:

Decomposition Algorithm

Step 0: Specify starting decisions

a. Decide on a summary measure.
b. Decide on a contrast.
c. Decide on the reference group for the mediator values.

Step 1: Estimate relationships in the data

a. Fit regression model(s) for the mediator(s) of interest with confounders of the mediator-outcome relationship as covariates.
b. Fit regression model(s) for the outcome with the mediator(s) of interest and same confounders as the mediator model.

Step 2: Form the Natural Course Pseudo-Population.

a. Use the coefficients from the mediator model(s) with the observed confounder values to simulate mediator values for each individual in the data.
b. Use the coefficients from the outcome model(s) together with the observed confounder values and the new simulated mediator values to simulate the outcome for each individual in the data. This is the natural-course pseudo-population.
c. Within this natural course pseudo-population, estimate the summary measure for both groups and then form the contrast of interest across groups.

Step 3: Form the Counterfactual Pseudo-Population

a. For the non-reference group, use the coefficients from the mediator model(s) with the observed confounder values to simulate counterfactual mediator values that follow the distribution of the mediator in the reference group.

b. Use the coefficients from the outcome model(s) together with the observed confounder values and simulated mediator values (counterfactual for the non-reference group and natural course for the reference group) to simulate the outcome for each individual in the data. This is the counterfactual pseudo-population.

c. Within this counterfactual pseudo-population, estimate the summary measure for both groups and then form the contrast of interest across groups.

Step 4: Compare the contrast of interest in the natural-course and counterfactual pseudo-populations.

To estimate standard errors and to produce stable estimates of the contribution, we have to address two types of variability. First, since we are drawing values of the mediators and outcomes from probability distributions, the exact values assigned to individuals can change across multiple draws. This results in the estimate of the contribution also changing across draws (known as Monte Carlo error). To reduce this error, we conduct Steps 2 and 3 multiple times, each time drawing a new set of mediator and outcome values. We then construct the contrasts for each draw and then average across all these draws to produce stable natural course and counterfactual estimates, before calculating the contribution in 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 before step 1, conducting the entire analysis k times, and then estimating the standard error of our decomposition estimates as the standard deviation of the estimates from the k bootstrap samples.

Our algorithm above treats the variables involved as time-fixed, which may not always be appropriate (5,8). The algorithm can be easily expanded, however, to allow for time-varying variables;

we present a time-varying version of the decomposition algorithm above in Appendix 2 based on Westreich et al. (2012) (26). A second important note is that the natural course is often used in g-formula analyses to validate the estimation models rather than as part of the estimand. In our algorithm, however, the natural course forms part of the contribution estimate. We use the natural course as the reference in the contribution rather than compare two counterfactual scenarios (such as one where all individuals smoke and one where no individuals smoke) so that the counterfactual scenarios are compared to the "as is" observed conditions. This approach is also advocated for by Hernán and Robbins (2016) as a more realistic comparison group for counterfactual analyses (27).

Both the size of and contribution of specific mediators to a health disparity are dependent on the scale that the disparity is measured on. For example, a difference in mortality between two populations and the contribution of smoking to this difference may vary based on whether the disparity is measured as a mortality risk ratio, a survival risk ratio, or an absolute difference in mortality rates. A major strength of our decomposition algorithm is that the researcher is not limited to one scale and can estimate and explain the disparity using multiple measures. This is because the decomposition algorithm works by first generating pseudo populations based around model-predicted values rather than by comparisons of model coefficients.

Empirical Example: Smoking's Contribution to Sex Differences in Life Expectancy in South Korea

We now demonstrate the application of the approach we outlined in the previous section to real data from the Korean Longitudinal Study of Aging. In the interest of providing a simple pedagogical example, we conduct a stylized analysis and thus the results should be interpreted cautiously. A more rigorous analysis that fully explores and accounts for the different sources of confounding and measurement error is outside the scope of this example, though the results of our example are in line with other literature on the contribution of smoking to sex differences in mortality (28).

Data: Korean Longitudinal Study of Aging

We use data from the 2006-2012 waves of the Korean Longitudinal Study of Aging, a nationally representative survey of South Korean individuals ages 45 and above (29). We use data on adults ages 50 and above from the baseline 2006 waves, using the subsequent waves for mortality follow-up. Our total sample consists of 7,615 individuals with 42,405 person-years of follow-up. We convert our data from a person to person-age format, with one observation for every age lived in the survey, along with a dichotomous indicator for whether an individual survived through or died on that age. Individuals leave the survey through death, censoring from loss to follow-up before 2012, or from censoring at the end of the survey period in 2012.

Main variables: outcome, mediator, and confounders

Our outcome of interest is dichotomous indicator for whether an individual died or survived to the next age and our primary mediator is a dichotomous indicator for whether an individual reported ever regularly smoking cigarettes. We adjust for the following potential confounders of the smoking-

mortality relationship: age, how frequently an individual reported drinking alcohol, schooling, urbanicity, and marital status.

Step 0: Specify a summary measure, contrast, and reference group

We consider two summary measures of mortality, the age-adjusted one-year risk of death (surviving to the next age) and period life expectancy at age 50. For the first summary measure, our contrast of interest is the risk ratio of mortality for men relative to women. We construct this contrast using the following Poisson regression on person-year observations (adjusting for age using indicator variables for five-year age groups):

$$log(E[Y|Female, Age]) = \alpha_0 + (\alpha_1 \cdot Female) + \sum_i (\alpha_i \cdot Agegr_i)$$

where α_1 is our estimate of interest. We use a Poisson regression here to just estimate the summary contrast (the exponent of α_1) but could have alternatively directly estimated an age-standardized risk ratio from the data. Importantly, because we are interested in the observed difference between men and women (adjusting for just age), we do not add any confounders to this model (19).

To construct period life expectancy, we first estimate age-specific mortality rates from the person-year data by dividing the number of deaths in each 5-year age group by the person-years of exposure in that same age group separately for men and women. Next, we convert these age-specific mortality rates into period life expectancies using standard life table techniques (30). Our contrast for this outcome is the absolute difference in life expectancy at age 50 between men and women.

For both summary measures and contrasts, we set the smoking levels among men to be equal to those among women as our counterfactual scenario.

Step 1: Estimate relationships in the data (using regression models)

Mediator model

We model the probability of ever regularly smoking for men and women using the following logistic regression model:

$$logit(E[Smk|Female, Age, C]) = \beta_0 + (\beta_1 \cdot Female) + (\beta_2 \cdot Age) + (\beta_3 \cdot Age \cdot Female) + \sum_i (\beta_{c_i} \cdot C_i)$$

Here, Smk is a binary variable for whether an individual self-reported ever regularly smoking, Sex is indicator variable for female, Age is continuous measurement of age, and C_i are the confounders described previously. We use this model to estimate the group -> causes association pathway in **Figure 1b**. We include the confounders in this model not to adjust for confounding but rather to allow us to predict and match the sex-specific smoking prevalence within confounder strata.

Outcome model

We model mortality as a function of smoking, sex, and the confounders by fitting the following logistic regression model:

$$logit(E[Y|Female, Smk, Age, C]) = \delta_0 + (\delta_1 \cdot Female) + (\delta_2 \cdot Smk) + (\delta_3 \cdot Female \cdot smk) + (\delta_4 \cdot Age) + (\delta_5 \cdot Age \cdot Female) + (\delta_6 \cdot Age \cdot Smk) + \sum_i (\delta_{c_i} \cdot C_i)$$

We use this model to estimate the causes -> outcome effect pathway in **Figure 1b**.

Steps 2 and 3: simulation to form the natural course and counterfactual pseudo-populations

Based on the results of the two models, we simulate the natural course and counterfactual pseudo-populations for both men and women. In **Figure 2,** we provide a step-by-step example of how to use

the regression estimates to form the simulated values for a single male individual in the data. The pseudocode in **Figure 3** and R code in the supplementary material demonstrate how to do this for all individuals in the data using common statistical software.

Steps 4: Calculate and compare the contrasts of interest and determine the percent contribution of smoking

Once pseudo-populations have been created, the final step is to calculate the contrasts of interest. We then estimate the contribution of smoking to sex differences in mortality by measuring how much the contrasts of the two summary measures changes between the natural course and counterfactual worlds. All steps needed to estimate the decomposition are also shown as pseudocode in **Figure 3**. We also provide code for how to estimate the example in R using our function *cfdecomp* in the supplementary material.

Results

Descriptive characteristics

Mean age was 66.2 for men and 67.4 for women (**Table 1**). A greater share of men was currently married compared to women (93% compared to 64%) due to a much higher proportion of widowhood among women (33% compared to 5%). There were important health and socioeconomic differences between men and women. Men were far more likely to smoke (61% compared to 4%) and drink regularly (proportion who reporting drinking at least once a week: 41% compared to 4%). Men were also substantially more likely to have completed more than middle school (46% compared to 17%).

Decomposition of the age-adjusted one-year risk of mortality

Men were 1.89 times (95% CI: 1.65, 2.14) more likely to die within one-year of an interview compared

to women (after adjusting for age) (**Table 2**). After setting men to have the same smoking distribution of women, this risk ratio reduced to 1.65 (95% CI: 1.38, 1.92). The resulting change corresponds to a (1 - 0.65/0.89) = 28% (95% CI: 0.08, 0.47) contribution of smoking to sex differences in the age-adjusted one-year risk of mortality.

Decomposition of period life expectancy at age 30

After converting the mortality risks into period life expectancy, we observe a large, 5.9-year difference (95% CI: -7.2, -4.4) in life expectancy at age 50 between men and women (**Table 3**). When we equalize levels of smoking between men and women, this difference in life expectancy reduces to just 4.3 years (95% CI: -5.7, -2.8), corresponding to a (1 - 4.3/5.9) = 27% (95% CI: 0.10, 0.44) contribution of smoking to sex-differences in adult life expectancy in South Korea.

Step 1: Regression estimates

$$logit(E[Y|Male, Smk, Age, C]) = -12.56 + (1.37 \cdot Male) + (0.12 \cdot Age) + (-0.01 \cdot Age \cdot Male) + (2.51 \cdot Smk) + (-0.30 \cdot Smk \cdot Male) + (-0.03 \cdot Age \cdot Smk) + (-0.17 \cdot Educ_1) + (-0.34 \cdot Educ_2) + (0.08 \cdot Rural) + (-0.64 \cdot Marit_1) + (-0.35 \cdot Marit_2) + (-0.50 \cdot Marit_3) + (-0.59 \cdot Alc_1) + (-0.37 \cdot Alc_2) + (-0.46 \cdot Alc_3) + (-0.26 \cdot Alc_4)$$

$$logit(E[Smk|Male, Age, C]) = -5.03 + (6.44 \cdot Male) + (0.03 \cdot Age) + (-0.04 \cdot Age \cdot Male) + (0.00 \cdot Educ_1) + (-0.35 \cdot Educ_2) + (0.01 \cdot Rural) + (-0.35 \cdot Marit_1) + (1.13 \cdot Marit_2) + (0.11 \cdot Marit_3) + (0.61 \cdot Alc_1) + (0.95 \cdot Alc_2) + (1.01 \cdot Alc_3) + (1.16 \cdot Alc_4)$$

Step 2: Simulate the natural course cohort

Covariate values for an example male in the data

$Male = 1$, $Age = 60$, $Educ = 2$, $Rural = 0$, $Marit = 3$, $Alc = 4$

Predict the probability of smoking

$$\hat{E}[Smk|Male = 1, Age = 60, Educ_1 = 0, Educ_2 = 1, Rural = 0, Marit_1 = 0, Marit_2 = 0, Marit_3 = 1, Alc_1 = 0, Alc_2 = 0, Alc_3 = 0, Alc_4 = 1] = \frac{e^{-5.03}e^{(6.44*1)}e^{(0.03*60)}e^{(-0.04*60*1)}e^{(-0.35*1)}e^{(0.11*1)}e^{(1.16*1)}}{1 + e^{-5.03}e^{(6.44*1)}e^{(0.03*60)}e^{(-0.04*60*1)}e^{(-0.35*1)}e^{(0.11*1)}e^{(1.16*1)}} = 0.85$$

Simulate the natural course smoking value

Smk = binomial.draw(probability = 0.85, draws = 1) = 1

Predict the probability of death

$$\hat{E}\left[Y\middle|\begin{matrix}Male = 1, Age = 60, Smk = 1, Educ_1 = 0, Educ_2 = 1, Rural = 0, Marit_1 = 0, \\ Marit_2 = 0, Marit_3 = 1, Alc_1 = 0, Alc_2 = 0, Alc_3 = 0, Alc_4 = 1\end{matrix}\right] = \frac{e^{-12.56}e^{(1.37*1)}e^{(0.12*60)}e^{(-0.01*60*1)}e^{(2.51*1)}e^{(-0.30*1*1)}e^{(-0.03*60*1)}e^{(-0.34*1)}e^{(-0.50*1)}e^{(-0.26*1)}}{1 + e^{-12.56}e^{(1.37*1)}e^{(0.12*60)}e^{(-0.01*60*1)}e^{(2.51*1)}e^{(-0.30*1*1)}e^{(-0.03*60*1)}e^{(-0.34*1)}e^{(-0.50*1)}e^{(-0.26*1)}} = 0.005$$

Simulate the natural course mortality value

Y = binomial.draw(probability = 0.005, draws = 1) = 0

Step 2: Simulate the counterfactual cohort (same male cohort member)

Predict the counterfactual probability of smoking using the same covariate values but Male = 0

$$\hat{E}[Smk|Male = 0, Age = 60, Educ_1 = 0, Educ_2 = 1, Rural = 0, Marit_1 = 0, Marit_2 = 0, Marit_3 = 1, Alc_1 = 0, Alc_2 = 0, Alc_3 = 0, Alc_4 = 1] = \frac{e^{-5.03}e^{(0.03*60)}e^{(-0.35*1)}e^{(0.11*1)}e^{(1.16*1)}}{1 + e^{-5.03}e^{(0.03*60)}e^{(-0.35*1)}e^{(0.11*1)}e^{(1.16*1)}} = 0.09$$

Simulate the counterfactual smoking value

Smk = binomial.draw(probability = 0.09, draws = 1) = 0

Predict the counterfactual probability of death using the same covariate values but Smk = 0

$$\hat{E}\left[Y\middle|\begin{matrix}Male = 1, Age = 60, Smk = 0, Educ_1 = 0, Educ_2 = 1, Rural = 0, Marit_1 = 0, \\ Marit_2 = 0, Marit_3 = 1, Alc_1 = 0, Alc_2 = 0, Alc_3 = 0, Alc_4 = 1\end{matrix}\right] = \frac{e^{-12.56}e^{(1.37*1)}e^{(0.12*60)}e^{(-0.01*60*1)}e^{(-0.34*1)}e^{(-0.50*1)}e^{(-0.26*1)}}{1 + e^{-12.56}e^{(1.37*1)}e^{(0.12*60)}e^{(-0.01*60*1)}e^{(-0.34*1)}e^{(-0.50*1)}e^{(-0.26*1)}} = 0.003$$

Simulate the counterfactual mortality value

Y = binomial.draw(probability = 0.003, draws = 1) = 0

Figure 2. A step-by-step example of how to use the regression estimates to form the simulated values for a single male individual in the data.

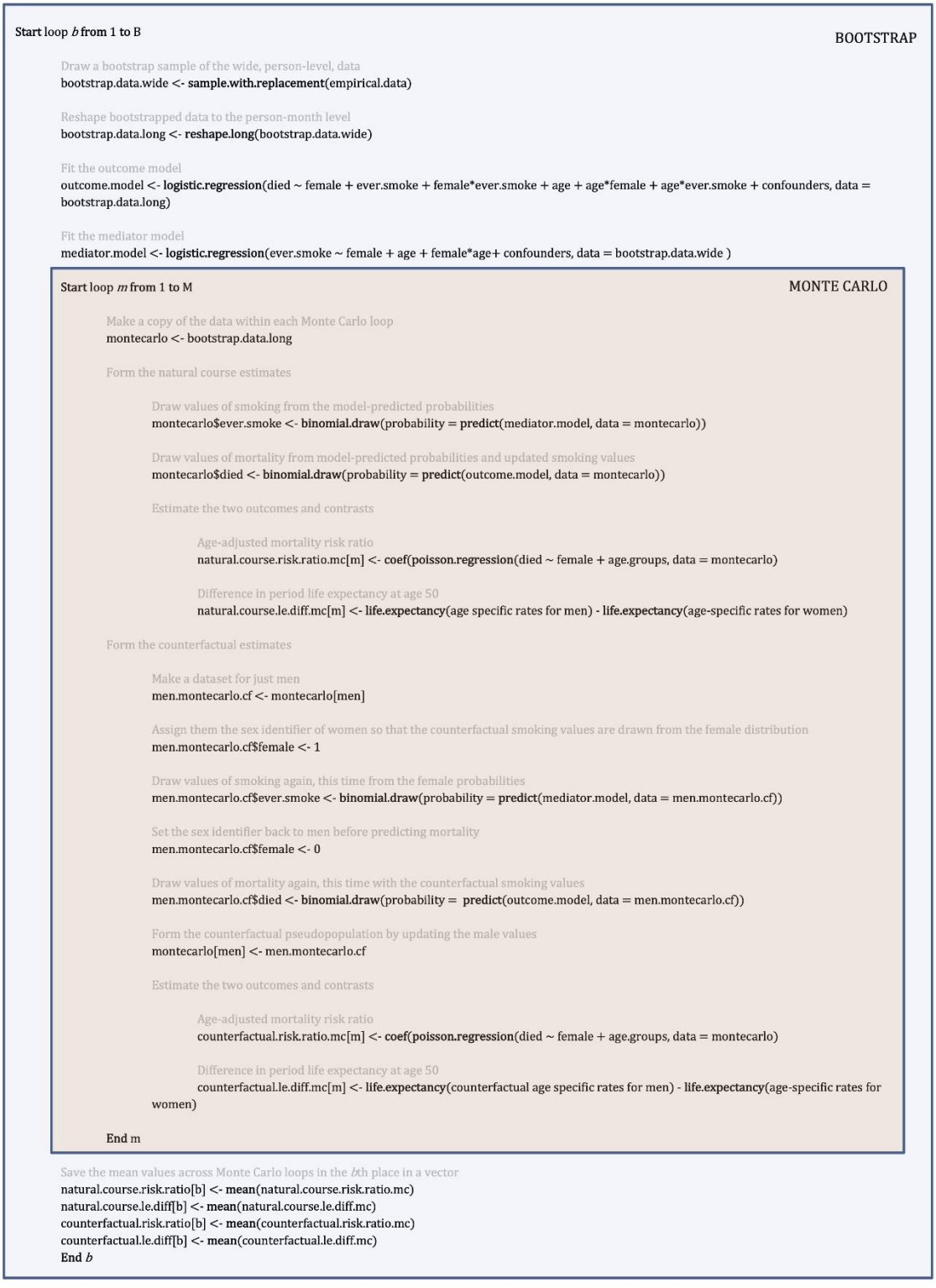

Figure 3. Pseudocode of a decomposition example.

Table 1 Descriptive characteristics of the sample at baseline, adults ages 50+, Korean Longitudinal Study of Aging, 2006.

	Men		Women	
	Mean	*SD*	*Mean*	*SD*
Age	66.2	9.0	67.4	9.9
	%	*N*	*%*	*N*
Marital status				
Never married	0.01	105	0.00	100
Married/partnered	0.93	17147	0.64	15350
Separated/divorced	0.02	349	0.02	499
Widowed	0.05	893	0.33	7962
Completed schooling				
None	0.09	1706	0.31	7299
Elementary or middle	0.45	8249	0.53	12574
More than middle	0.46	8539	0.17	4038
Rural	0.27	4987	0.27	6534
Ever smoker	0.61	11276	0.04	1015
Alcohol consumption				
None/less than once a month	0.43	7868	0.87	20808
One to several times a month	0.16	3040	0.08	2000
One to several times a week	0.28	5119	0.04	906
Most days of the week	0.05	935	0.00	113
Every day of the week	0.08	1532	0.00	84

Table 2 Estimates of the contribution of smoking to the age-adjusted one-year mortality risk ratio using the counterfactual decomposition method, Korean Longitudinal Study of Aging, 2006-2012.

	Natural course RR (95% CI)	Counterfactual RR (95% CI)	Percent contribution (95% CI)
Mortality risk ratio for men relative to women	1.89 (1.65, 2.14)	1.65 (1.38, 1.92)	28% (8%, 47%)

Table 3 Estimates of the contribution of smoking to sex differences in period life expectancy at age 50 (e_{50}) using the counterfactual decomposition method, Korean Longitudinal Study of Aging, 2006-2012.

	Natural Course e_{50}	Natural Course Δ	Counterfactual e_{50}	Counterfactual Δ	Percent contribution
Women	36.8				
(95% CI)	(35.9, 37.7)				
Men	30.9	-5.9	32.5	-4.3	27%
(95% CI)	(30.0, 31.8)	(-7.2, -4.4)	(31.3, 33.7)	(-5.7, -2.8)	(10%, 44%)

Discussion

We introduce a general yet easily applied procedure for implementing counterfactual decompositions using the parametric g-formula and Monte Carlo integration (19). We demonstrate this approach by estimating the contribution of smoking to sex differences in mortality in South Korea. We first decompose the simple contrast of the age-adjusted mortality risk ratio for men relative to women and then demonstrate how to decompose functions of population risks by decomposing the sex difference in period life expectancy at age 50. We find that the large smoking difference between men and women in South Korea explains 27-28% of the age-adjusted mortality risk ratio and sex difference in life expectancy at age 50.

The age-adjusted mortality risk could also be decomposed using closed-form decomposition equations (12,13,19). The algorithm we outline does not replace closed-form decomposition approaches but rather provides an alternative using simulations, which provides three main advantages. First, we can decompose summary measures based on any outcome distribution in the GLM family without having to derive or use separate decomposition equations depending on whether an outcome is binomially, Poisson, or normally distributed. Moving between outcome distributions simply requires changing the regression type used to model the outcome in the decomposition algorithm.

The second advantage of the simulation algorithm is that we can easily switch between different contrasts since we effectively re-generated entire micro-populations for the observed and counterfactual worlds. For example, once natural course and counterfactual pseudo-populations have been generated, we decomposed the risk ratio by estimating Poisson regressions of mortality on sex within both pseudo-populations and measuring how the risk ratio changes between the natural course and counterfactual worlds. If we were instead interested in decomposing the odds ratio, we would simply switch from Poisson to logistic regressions and compare the odds ratios. This pseudo-

population perspective is powerful because it easily allows for comparisons of any contrast we can think of. Indeed, the third advantage of the simulation approach to decomposition is that we can decompose summary measures that are based around complex functions of population means and proportions. Period life expectancy is an example of such a summary measure since it is a function of age-specific mortality risks.

Despite these advantages, our algorithm comes with important trade-offs compared to existing decomposition implementations. Compared to the closed-form equations our approach requires substantial computational power and time. This is not a trivial consideration and decompositions with large datasets may take hours to even days to complete even when considerable computational power is available. Furthermore, as with any method seeking to provide causal explanations, the causal validity of the decomposition results hinges on assumptions of exchangeability (also known as no unmeasured confounding), common support (positivity), and consistency. We discuss these three issues in more detail in Appendix 1 for interested readers.

Conclusions

Decomposing the sources of differences in health and other outcomes is a key research endeavor in epidemiology and other population health sciences. We describe an implementation of the counterfactual decomposition that builds on and generalizes the rich existing body of work on decomposition methods in the health and social sciences. The approach provides a highly flexible and easily implemented way of estimating decompositions that are grounded in potential outcomes and counterfactual theory and applicable to a wide range of population health questions.

References

1. Peyvandi S, Baer RJ, Moon-Grady AJ, Oltman SP, Chambers CD, Norton ME, et al. Socioeconomic mediators of racial and ethnic disparities in congenital heart disease outcomes: a population-based study in California. J Am Heart Assoc. 2018;7(20):e010342.

2. Sudharsanan N, Ho JY. Rural–urban differences in adult life expectancy in Indonesia: a parametric g-formula based decomposition approach. Epidemiology. 2020;

3. Martikainen P, Mäkelä P, Peltonen R, Myrskylä M. Income Differences in Life Expectancy: The Changing Contribution of Harmful Consumption of Alcohol and Smoking. Epidemiology. 2014 Mar;25(2):182–90.

4. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. Psychol Methods. 2010;15(4):309–34.

5. Lin S-H, Young J, Logan R, Tchetgen EJT, VanderWeele TJ. Parametric mediational g-formula approach to mediation analysis with time-varying exposures, mediators, and confounders. Epidemiol Camb Mass. 2017;28(2):266.

6. Lin S-H, Young JG, Logan R, VanderWeele TJ. Mediation analysis for a survival outcome with time-varying exposures, mediators, and confounders. Stat Med. 2017;36(26):4153–4166.

7. De Stavola BL, Daniel RM, Ploubidis GB, Micali N. Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. Am J Epidemiol. 2015;181(1):64–80.

8. Vansteelandt S, Daniel RM. Interventional effects for mediation analysis with multiple mediators. Epidemiol Camb Mass. 2017;28(2):258.

9. VanderWeele TJ, Robinson WR. On causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiol Camb Mass. 2014;25(4):473.

10. Blinder AS. Wage discrimination: reduced form and structural estimates. J Hum Resour. 1973;436–455.

11. Oaxaca R. Male-female wage differentials in urban labor markets. Int Econ Rev. 1973;693–709.

12. Powers DA, Yun M-S. 7. Multivariate Decomposition for Hazard Rate Models. Sociol Methodol. 2009;39(1):233–263.

13. Yun M-S. Decomposing differences in the first moment. Econ Lett. 2004;82(2):275–280.

14. Machado JA, Mata J. Counterfactual decomposition of changes in wage distributions using quantile regression. J Appl Econom. 2005;20(4):445–465.

15. Kitagawa EM. Components of a difference between two rates. J Am Stat Assoc. 1955;50(272):1168–1194.

16. Arriaga EE. Measuring and explaining the change in life expectancies. Demography. 1984;21(1):83–96.

17. Horiuchi S, Wilmoth JR, Pletcher SD. A decomposition method based on a model of continuous change. Demography. 2008;45(4):785–801.

18. Andreev EM, Shkolnikov VM, Begun AZ. 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. Demogr Res. 2002;7:499–522.

19. Jackson JW, VanderWeele TJ. Decomposition analysis to identify intervention targets for reducing disparities. Epidemiology. 2018;29(6):825–835.

20. Nandi A, Glymour MM, Subramanian S. Association among socioeconomic status, health behaviors, and all-cause mortality in the United States. Epidemiology. 2014;25(2):170–177.

21. Bijlsma M, Sudharsanan N, Li P. cfdecomp: Counterfactual Decomposition: MC Integration of the G-Formula [Internet]. 2020. Available from: https://cran.r-project.org/package=cfdecomp

22. Bijlsma MJ, Wilson B. Modelling the socio-economic determinants of fertility: a mediation analysis using the parametric g-formula. J R Stat Soc Ser A Stat Soc. 2019;

23. VanderWeele TJ, Vansteelandt S, Robins JM. Effect decomposition in the presence of an exposure-induced mediator-outcome confounder. Epidemiol Camb Mass. 2014;25(2):300.

24. Young JG, Tchetgen Tchetgen EJ. Simulation from a known Cox MSM using standard parametric models for the g-formula. Stat Med. 2014;33(6):1001–1014.

25. 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.

26. Westreich D, Cole SR, Young JG, Palella F, Tien PC, Kingsley L, et al. The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. Stat Med. 2012;31(18):2000–2009.

27. 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.

28. Preston SH, Wang H. Sex mortality differences in the United States: The role of cohort smoking patterns. Demography. 2006;43(4):631–646.

29. Jang S-N. Korean Longitudinal Study of Ageing (KLoSA): overview of research design and contents. Encycl Geropsychology. 2015;1–9.

30. Preston SH, Heuveline P, Guillot M. Demography measuring and modeling population processes. 2000.

Appendix 1: Causal Inference Considerations

Exchangeability

Results from our approach are only valid if the underlying assumption of no unmeasured confounding of the mediator-outcome relationship is correct (exchangeability). Even with a large number of confounders, this is a strong assumption and thus the results need to be interpreted cautiously with consideration to the magnitude of bias that potential unmeasured confounders may introduce. For example, smoking is likely correlated with other unhealthy behaviors that also affect mortality that we were not able to adjust for. Therefore, when we "intervene" on smoking to produce our decomposition estimate, we are also change the other unobserved behaviors correlated with smoking. Thus, our decomposition estimate of the contribution of smoking to sex differences in life expectancy is likely an overestimate. Bias analyses may be a promising way to evaluate the causal validity of the decomposition estimates.[1–3] Importantly, since we are interested in the observed difference between social groups, we deliberately do not equalize levels of the confounders of the social group-outcome relationship across social groups.

Positivity

One conceptual issue that may arise is a lack of common support (also known as positivity) of the mediator distribution across groups. For example, suppose we are interested in equalizing the distribution of smoking between men and women with very high levels of schooling. If the low SES group has total schooling 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 schooling group would respond to having schooling values above 9. In such a case, one may be forced to assume that the relationship between total schooling values above 9 in the low SES group is the same as that of the high SES group or be willing to extrapolate the model estimates outside the range of observed data.

Consistency

An important issue is whether "interventions" to equalize levels of the mediators satisfy the consistency assumption and thus can be justified as causal effects - a prerequisite for providing the decomposition results a causal interpretation. For example, we estimate the contribution of smoking to sex differences in life expectancy using a measure of ever smoking; therefore, the change in mortality produced by a real smoking cessation intervention may not be well approximated by a contrast of ever and never smokers. In this case, our example is better conceptualized as a thought experiment that asks what if the share of men who ever started smoking (and survived beyond age 30) never exceeded the share of ever smokers among women?

Appendix 2: Decomposition with time-varying variables

When the process being investigated involves time-varying variables, including potentially time-varying confounding variables or intermediate confounders, the algorithm to perform decomposition becomes more involved.[4] We here denote the steps that could be taken in a cross-lagged model. Note that since this describes a decomposition model, the exposure variable (group) is time-fixed and hence is not estimated. In a mediation analysis -- rather than a decomposition -- with a time-varying exposure, the exposure should also be modelled and simulated if a 'natural course scenario' (a replication of the empirical data, often used for validation purposes) is required.

Decomposition Algorithm for Time-varying Covariates

Step 0: Specify starting decisions

d. Decide on a summary measure.
e. Decide on a contrast.
f. Decide on the reference group for the mediator values.

Step 1: Estimate relationships in the data

Fit regression model(s) for the time-varying variables of interest (time-varying confounders, mediators, and outcomes) with confounders of the mediator-outcome relationship as covariates. For time-varying variables that are independent of the measured covariates, a model with time itself (as a categorical variable, or e.g. modelled with splines) as a covariate, and potentially with baseline covariates, could be used. Note that separate models can be fitted per group, or interactions with a group identifier could be used. An example specification of a cross-lagged model could be:

$$f(E[M_{t+1}|C_t, X_t, Y_t]) = \beta_0 + \beta_1 \cdot C_t + \beta_2 \cdot C + \beta_3 \cdot X_t + Y_t$$

Where f is an appropriate link function, the index refers to time, M refers to a mediator of interest, C to time-varying confounders, X to other time-varying mediators whose joint contribution we are

interested in, and Y to a time-varying outcome variable that is also allowed to affect future values of the mediator(s).

Step 2: Form the Natural Course Pseudo-Population.

d. Take observed values from the empirical data at t=1.

e. Using the observed values at t=1, use the models for the time-varying variables to simulate values at t=2.

f. Akin to step 2a, continue with taking simulated values at *t* to simulate values at *t+1* until the end of follow-up. This is the natural course pseudo-population.

g. Within this natural course pseudo-population, estimate the summary measure for both groups and then form the contrast of interest across groups.

Step 3: Form the Counterfactual Pseudo-Population.

a. Take observed values from the empirical data at t=1. For the non-references group(s), draw the mediator values from the distribution of the reference group (see Wang & Arah 2015 for the difference between the controlled direct effect and the stochastic controlled direct effect).[5] If the distribution of the mediators in the reference group changes over time, this should be taken into account.

b. Using the observed (and now partially altered) values at t=1, use the models for the time-varying variables to simulate values at t=2. For the non-reference group(s), simulate mediator values that follow the distribution of the reference group.

c. Akin to step 2a, continue with taking simulated values at *t* to simulate values at *t+1* until the end of follow-up, and for the non-reference groups continue simulating mediator values that follow the distribution of the reference group. This is the counterfactual pseudo-population.

d. Within this counterfactual pseudo-population, estimate the summary measure for both groups and then form the contrast of interest across groups.

Step 4: Compare the contrast of interest in the natural-course and counterfactual pseudo-populations.

To estimate standard errors and to produce stable estimates of the contribution, we have to address two types of variability. First, since we are drawing values of the mediators and outcomes from probability distributions, the exact values assigned to individuals can change across multiple draws. This results in the estimate of the contribution also changing across draws (known as Monte Carlo error). To reduce this error, we conduct Steps 2 and 3 multiple times, each time drawing a new set of mediator and outcome values. We then construct the contrasts for each draw and then average across all these draws to produce stable natural course and counterfactual estimates, before calculating the contribution in 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 before step 1, conducting the entire analysis k times, and then estimating the standard error of our decomposition estimates as the standard deviation of the estimates from the k bootstrap samples.

Appendix 3: Formal counterfactual approach

Our approach requires that we estimate what an outcome (Y) would have been among one group (group B) if they were set to have the same distribution of the mediator (M) as another group (group A). We first define the potential outcome for an individual when the mediator M is set to a specific value m as $Y(M = m)$. We denote the distribution of M in group A as f_M^A and that in group B as f_M^B and the potential outcome for an individual when the mediator is set to a value drawn from this distribution as $Y(M{\sim}f_M^A)$ and $Y(M{\sim}f_M^B)$, respectively. Equalizing M as described , we are now interested in the value of the outcome (Y) for individuals in group B when the mediator (M) is redistributed to f_M^A: $Y^B(M{\sim}f_M^A)$).

Next, we need to formally define our summary measure and population contrast of interest. For this exposition, we will use the mean of Y as our summary measure and the difference in this mean between groups A and B, $E[Y^A] - E[Y^B]$, as our contrast. Given this summary measure and contrast, we are now interested in the mean difference in the outcome between groups when the mediator (M) among group B has been redistributed to f_M^A: $E[Y^A] - E[Y^B(M{\sim}f_M^A)]$. The second term is the counterfactual potential outcome since it is not directly observable in the data. One way to reveal how to estimate this quantity is by expanding the observed mean outcome among group B by conditioning on the different values of the mediator (M) found in f_M^B:

$$E[Y^B] = \sum_{m \in f_M^B} E[M = m, B] \cdot P(B) \qquad (2)$$

Within this expression, the distribution of the mediator (M) for group B, f_M^B, is captured by the set of probabilities, $P(M = m|B)$, for each value of m found in f_M^B. Therefore, if we wanted to estimate what the expected value of Y^B would be if group B had the same distribution of the mediator as group A ($E[Y^B(M{\sim}f_M^A)]$), we could replace the probabilities of observing each value of M in group B with

the corresponding probability of observing that value in group A ($P(M = m|A)$). Then we would estimate the potential outcome as:

$$E[Y^B(M{\sim}f_M^A)] = \sum_{m \in f_A()} E[M = m, B] \cdot P(A) \quad (3)$$

This is simply a direct standardization within confounder strata.

Unfortunately, in most observational research, this approach will not lead to a correct estimate of the counterfactual average potential outcome since it assumes that the expected value of the outcome Y when M is set to a specific value m_i among those with $m \neq m_i$ can be estimated as the observed expected value for those with $m = m_i$. This condition, known as exchangeability,[6] is often a strong assumption given that there are likely other systematic ways those with different values of M differ that would affect their value of Y. Therefore, in the presence of confounding variables (C), $E[Y^B(M{\sim}f_M^A)] \neq \sum_{m \in f_M^A} E[Y|M = m, B] \cdot P(M = m|A)$. However, this equality will hold within strata of C:

$$E[C = c] = \sum_{m \in f_M^A} E[M = m, C = c, B] \cdot P(C = c, A) \quad (4)$$

This is because within strata, there is no difference in the value of the confounders between those with different levels of the mediator. Therefore, differences in stratum-specific potential outcomes are not confounding the effect of the mediator with the effect of different confounder values.

We can now estimate $E[Y^B(M{\sim}f_M^A)]$ by aggregating these conditional potential outcome estimates across the strata of C and M:

$$E[Y^B(M{\sim}f_M^A)] = \sum_C \sum_{f_M^A} E[Y|M = m, C = c, B] \cdot P(M = m|C = c, A) \cdot P(C = c, B) \quad (5)$$

Estimating this equation amounts to first stratifying by all values of C. Next, within each of these strata, estimating what the outcome for individuals in group B would be if their mediator values were re-distributed to the mediator distribution of group A in that same confounder stratum. To do this, we would estimate the expected value of the outcome for group B individuals for each value of the mediator found in the mediator distribution of group A individuals in that same confounder stratum $f^{A}_{M|C=c}$. We would then multiply these stratum-specific counterfactual-expected outcome values by the share of the stratum with that specific value of the mediator in group A ($P(M = m|C = c, A)$) and then sum across strata of M and C. This second step matches the distribution between groups by equalizing the share of individuals with each value of m in group B to that share in group B (within confounder strata).

At this point, estimating the decomposition first defined in Eq. 1.requires the following three quantities: $E[Y^A]$, t $E[Y^B]$, and $E[Y^B(M{\sim}f^A_M)]$. Inserting these quantities into Eq. (1) leads to our analytic expression for the contribution of the mediator M to group differences in the outcome Y:

$$Contribution = 1 - \frac{E[Y^B(M{\sim}f^A_M)] - E[Y^A]}{E[Y^B] - E[Y^A]} \quad (6)$$

References of Appendix 3

1. Carnegie, N. B., Harada, M. & Hill, J. L. Assessing sensitivity to unmeasured confounding using a simulated potential confounder. *J. Res. Educ. Eff.* **9**, 395–420 (2016).

2. VanderWeele, T. J. & Arah, O. A. Bias formulas for sensitivity analysis of unmeasured confounding for general outcomes, treatments, and confounders. *Epidemiol. Camb. Mass* **22**, 42–52 (2011).

3. Bijlsma, M. J. & Wilson, B. Modelling the socio-economic determinants of fertility: a mediation analysis using the parametric g-formula. *J. R. Stat. Soc. Ser. A Stat. Soc.* (2019).

4. Westreich, D. *et al.* The parametric g-formula to estimate the effect of highly active antiretroviral therapy on incident AIDS or death. *Stat. Med.* **31**, 2000–2009 (2012).

5. Wang, A. & Arah, O. A. G-computation demonstration in causal mediation analysis. *Eur. J. Epidemiol.* **30**, 1119–1127 (2015).

6. Greenland, S. & Robins, J. M. Identifiability, exchangeability, and epidemiological confounding. *Int. J. Epidemiol.* **15**, 413–419 (1986).