APPENDIX A. Parametric g-computation and total effect decomposition.

G-formula for causal mediation analysis

Conventional regression, controlling for time-varying characteristics, assumes that there exists no exposure-induced mediator-outcome confounding; in other words, that the cross-world independence assumption holds (Vanderweele 2014; Wodtke and Zhou 2020). The counterfactual comparison is between two populations that vary on exposure status (in this case, being racialized as Black or white in a system of racism), assuming *nothing else* changes across the two populations over time. As described above, previous research has demonstrated this assumption is untenable in the quantitative study of racism as a social process, where being racialized within a system of racism affects virtually all other observed variables over time (Sen and Wasow 2016).

The "g-formula" or "g-computation" is a generalization of standardization that allows for the estimation of unconfounded summary effects in the presence of observed post-treatment confounding. The g-formula was developed in the formal quantitative causal inference literature and is flexible for estimating any counterfactual contrast (A. Naimi, Cole, and Kennedy 2016; Robins 1986; Wang and Arah 2015). Equation 1 illustrates the population mean health outcome, $E[Y]$, standardized across all values of an exposure variable, X (e.g., being racialized as Black within the system of racism governing the distribution of Y).

$$E[Y] = \sum_{x} P(Y = y|X = x)\, P(X = x) \tag{1}$$

This generalized formula, or "g-formula," for the mean outcome at a given age can be extended over all stratifying variables, $\boldsymbol{V}$, which confound the association between X and Y, as well as

variables which mediate the association, $\boldsymbol{M}$ (i.e., fall along the causal pathway). We use $P(y|x)$ as shorthand for $P(Y = y|X = x)$, $P(\boldsymbol{v})$ as shorthand for $P(\boldsymbol{V} = \boldsymbol{v})$, etc. In Equation 2, we illustrate the g-formula for the expectation of Y given exposure level $X = x$.

$$E[Y^x] = \sum_m P(Y|x, m, \boldsymbol{v})\, P(m|x, \boldsymbol{v})\, P(\boldsymbol{v}) \tag{2}$$

We extend Equation 2 to consider a mediator M which is dependent on a previous mediator L. We use the following example in Equation 3:

- $Y =$ normalized z-score on cardio-metabolic risk index (continuous).
- $X =$ self-identified race (x = white, x^* = Black).
- $M =$ second mediator (e.g., college attainment) (m = index value, m^* = reference value).
- $L =$ first mediator (e.g., parent college attainment) (l = index value, l^* = reference value).
- $\boldsymbol{V} =$ vector of 1) exposure-outcome confounders and 2) mediator-outcome confounders not influenced by exposure. This only includes age and gender, but we clarify in the Discussion how these may also be treated as mediators.

$$E[Y^x] = \sum_m \sum_l P(Y|x, m, l, \boldsymbol{v})\, P(m|x, l, \boldsymbol{v})\, P(l|x, \boldsymbol{v})\, P(\boldsymbol{v}) \tag{3}$$

When considering the effect on Y of changes to X via a specific mediator M, variables such as L are often referred to as "exposure-induced mediator-outcome confounders" because they are affected by the exposure and confound the relationship between M and Y (A. Naimi et al. 2016; Wang and Arah 2015). The presence of such confounding means that we cannot estimate the counterfactual associated with a given value of X while holding L constant (as in conventional regression or matching estimators, e.g., Baron-Kenny mediation), because such a world would be

impossible to observe. In terms of our theoretical framework, this reflects on the critiques discussed above: what would it mean to consider the effect of being racialized one way vs. another *without* anything else changing? At best, this conventional approach involves describing a marginal counterfactual that is difficult to interpret because changing this exposure requires considering changes in everything else that is influenced by and acts on racialized status. At worst, this approach reifies the notion that race is a construct that can be considered separately and independently from other factors such as socioeconomic status (Kohler-Hausmann 2019; Sen and Wasow 2016; Zuberi and Bonilla-Silva 2008).

The generalization of the entire conditional probability space in Equation 3 is the critical contribution of the g-formula standardization because it has important implications for estimating population-level counterfactuals without requiring that all variables be fixed at their means or reference values. Rather, in decomposing any population disparity in Y by exposure X (or the total "average treatment effect" of X on Y), specific mediating effects can be considered while other variables that are in any way dependent on the exposure take on the values they *would have had* under that particular counterfactual exposure history (Daniel et al. 2015; A. I. Naimi et al. 2016; Naimi 2016; Wang and Arah 2015). In conventional regression models (e.g., Baron-Kenny mediation) or demographic decomposition (e.g., Das Gupta or Kitagawa decomposition), estimates of counterfactual change are calculated under the assumption that no other conditional probabilities change as a result of the exposure changing (Sudharsanan and Bijlsma 2022). In contrast, g-formula standardization makes explicit the sum of all "cascades" of conditional probabilities for all variables as the cohort ages through that time and space. The conditional probabilities of all mediators (M, L) in Equation 3 can be expanded to include the

specific dependence structure for each variable as described by a given causal model (in this analysis, the DAG in Figure 2).

Total effects calculated via g-computation are analogous to effects obtained by marginal structural models (Lee and Jackson 2017; Lin et al. 2017; Robins, Hernán, and Brumback 2000; Wodtke, Harding, and Elwert 2011). However, the g-formula provides an intuitive method for decomposing this total effect or disparity into additive direct, interactive, and indirect pathways of accumulation via predicting and differencing counterfactual quantities rather than relying on often complex weighting schemes. In the simplest case of a single mediator *M* (treating the first mediator in Equation 3, *L*, as a post-treatment confounder) the difference between $E[Y|x]$ and $E[Y|x^*]$ can be decomposed into the controlled direct effect (CDE; *racism via unobserved mediating pathways*), the proportion attributable to interaction via each mediator M (PAI; *racial discrimination in the underlying system connecting M to Y*), and the pure indirect effect via each mediator M (PIE; *emergent discrimination*) (Reskin 2012; VanderWeele 2014; Wang and Arah 2015).

- $\text{CDE}_{M=m^*} = E[Y_{xm^*}] - E[Y_{x^*m^*}]$
- $\text{PAI}^{(M)} = E[(Y_{xm} - Y_{x^*m} - Y_{xm^*} + Y_{x^*m^*})(M_x)]$
- $\text{PIE}^{(M)} = E[Y_{x^*M_x}] - E[Y_{x^*M_{x^*}}]$

The case of multiple sequential mediators

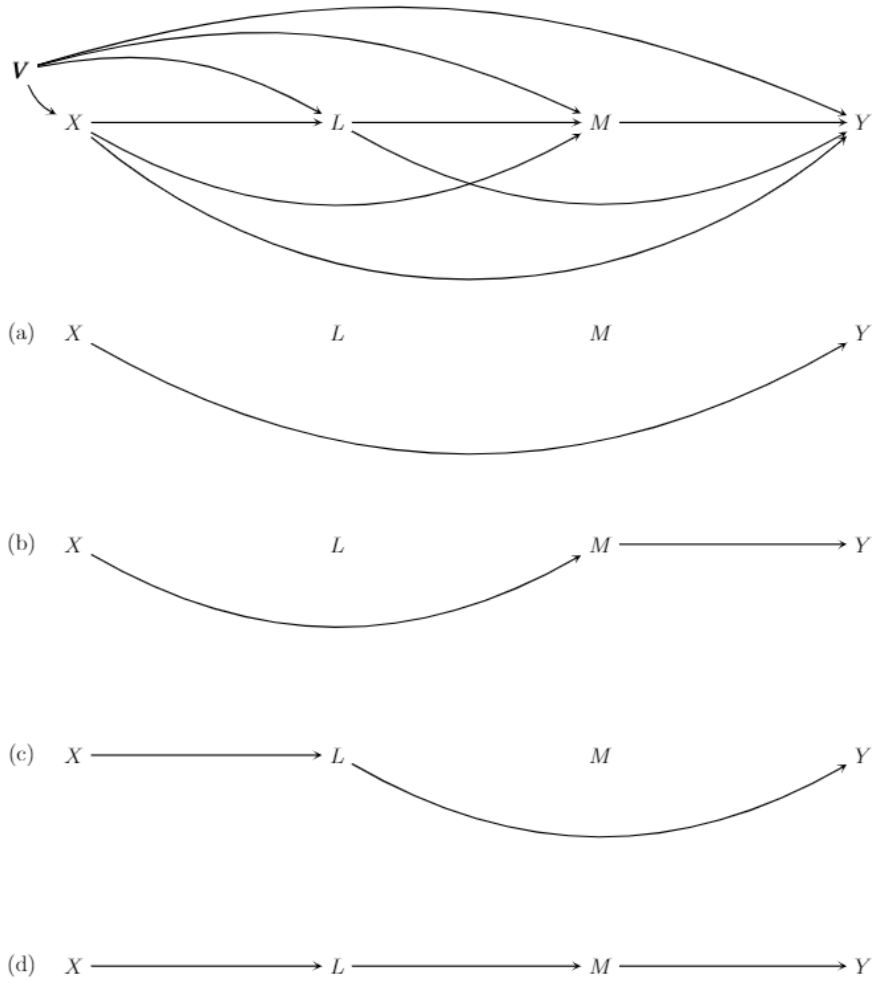
Wang & Arah (2015) describe the general application of g-computation to causal mediation and effect decomposition considering one mediator with an exposure interaction. There have been many useful extensions developed in the context of multiple dependent mediators, for example: Daniel et al. (2015) (considering multiple dependent mediators, but no

exposure interactions), Shi et al. (2021) (considering a joint set of multiple dependent mediators), and Zhou & Yamamoto (2020) (considering path-specific effects via multiple dependent mediators) (Daniel et al. 2015; Shi, Choirat, and Valeri 2021; Zhou and Yamamoto 2020).

In the present study, we use a relatively straightforward extension of the decomposition in Wang & Arah (2015) to describe separate mediated effects (PAI, PIE) via multiple sequential mediators. As a simple example of two mediators as described in Equation 3 (L = parent college attainment, M = college attainment), we decompose the total average treatment effect into the following:

- $\text{CDE}_{L=l^*,M=m^*} = E[Y_{xl^*m^*}] - E[Y_{x^*l^*m^*}]$
- $\text{PAI}^{(L)} = E[(Y_{xlm^*} - Y_{x^*lm^*} - Y_{xl^*m^*} + Y_{x^*l^*m^*})(L_x)]$
- $\text{PIE}^{(L)} = E[Y_{x^*L_xM_{x^*l^*}}] - E[Y_{x^*L_{x^*}M_{x^*l^*}}]$
- $\text{PAI}^{(M)} = E[(Y_{xl^*m} - Y_{x^*l^*m} - Y_{xl^*m^*} + Y_{x^*l^*m^*})(M_{xl})]$
- $\text{PIE}^{(M)} = E[Y_{x^*L_{x^*}M_{xl}}] - E[Y_{x^*L_{x^*}M_{x^*l^*}}]$

We note several differences in this decomposition compared to the single-mediator case. In discussing these differences and comparing this effect decomposition with other estimands and methods, consider the plot below for the two-mediator case, adapted from Figure 1 in Zhou & Yamamoto (2022):



First, the CDE in the two-mediator case is evaluated at the reference value for both mediators (e.g., no college degree *and* parent with no college degree); pathway (a) above.

Second, pathway (d) is now removed from the PAI/PIE of the most "upstream" mediator, *L*. In removing pathway (d) from these effects and only focusing on pathway (c), calculation of the PAI/PIE via *L* does not change depending on the distribution of *M*; using the reference value in the equations above is a convenient way to avoid picking up the PAI via *M* within these effects via *L*, as this calculation includes counterfactual values under different levels of the exposure. In other words, we are attempting to isolate change in *Y* that can be attributed to the *racialized returns of a particular mediator*. In this example, the PAI via *L* thus refers to the

change in *Y* that would be expected if individuals were *racialized as Black by the system governing the distribution of* ***parental education*** *and its* ***relation*** *to cardiometabolic risk* but racialized as white by the system governing the relation between their own education and cardiometabolic risk. Conversely, the PAI via *M* refers to the change in *Y* that would be expected if individuals were *racialized as Black by the system governing the distribution of their* ***own education*** *and its* ***relation*** *to cardiometabolic risk* but racialized as white by the system governing the relation between parental education and cardiometabolic risk.

Third, in calculating the PAI/PIE via *M*, the distribution of *M* is equal to its counterfactual distribution under treatment (M_{xl}), which includes the ways in exposure influences *L* which in turn influences *M;* pathways (b) and (d) illustrated above. In other words, our decomposition somewhat preferences "downstream" mediators to the extent that the PIEs are relatively important in the effect decomposition.

To summarize, the conceptual ambiguity of separating mediation effects given multiple mediators is in large part 1) how to handle pathway (d) above and 2) the separability of the PAI via multiple pathways. Our decomposition is effectively putting pathway (d) into the *mediated effects via M*; all separate mediated effects (PIE/PAI via *L*, PIE/PAI via *M*) then add up to the joint mediated effects via *L/M* (as in Shi et al. 2021). This particular decomposition, which can be extended to any number of sequential mediators, might be more or less theoretically justified depending on the research question. Users seeking to do their own analyses with multiple sequential mediators should consider whether this implementation is aligned to their specific theoretical estimand(s) (Lundberg, Johnson, and Stewart 2021), compared to the joint mediated effect of the set of multiple sequential mediators (*CMAverse* package) or individual path-specific

effects (*paths* package); see Zhou & Yamamoto (2022) for a comprehensive discussion on this topic.

Estimation

G-computation requires estimating multiple counterfactual values for Y using the relevant g-formula (e.g., Equation 3). This is achieved via simulation, which involves the following general steps (Lin SH et al. 2017; Shi et al. 2021). First, we fit a survey-weighted generalized linear model for each mediator (e.g., *L* and *M*) and the outcome (*Y*) using the *survey* package in R. These models take the general form indexed by individual (*i*).:

$$Y_i = \beta_0 + \boldsymbol{\beta_1}(\boldsymbol{M}_i * X_i) + \boldsymbol{\beta_2}(\boldsymbol{V}_i) + \varepsilon_i$$

Models parameterize each mediator and the outcome with appropriate likelihoods (normal, binomial) and link functions (identity, logit) given the structure of each dependent variable (e.g., *L, M, Y*). All models include age and gender ($\boldsymbol{V}_i$), account for the survey design of Add Health by including longitudinal survey weights, and standard errors are clustered by individual using the *survey* R package (Lumley 2010, 2018). Coefficient estimates from all generalized linear models are reported in Appendix Table A.1.

Given these fitted models, we then repeat the decomposition below 1000 times to propagate uncertainty to final effect estimates, which are summarized by the means and 95% intervals of those 1000 estimates:

1. Draw a random multivariate-normal sample of parameters from all survey-weighted models using fitted coefficients and variance-covariance matrices.
2. Create 30 replicates of the dataset to remove Monte Carlo error arising from stochastic individual-level response prediction in estimating counterfactuals.

3. Simulate two counterfactual datasets (one under exposure x and one under exposure x^*) by predicting forward with the sample of parameter estimates (i.e., What if this cohort had all been racialized as white by the observed mediating systems?).
4. Use values drawn from these simulations to calculate all necessary population-average counterfactual quantities for effect decomposition (e.g., $E[Y_{x^*L_xM_{x^*l^*}}]$). Calculate the CDE and the PAI/PIE for each mediator by differencing these quantities as in the equations described above.

Two types of non-random missing data may result in biased effect estimates: item-nonresponse and censoring of observations. To avoid bias resulting from item-nonresponse, we create 30 multiply imputed datasets using chained equations to account for missing observations (Buuren and Groothuis-Oudshoorn 2011). To avoid bias resulting from the censoring of observations, the g-computation simulations begin with the full sample in the first wave and every individual is simulated through all subsequent waves. In effect calculations, we are then including all simulated person-years that were censored in the survey sample, assuming these individuals would have responded the same in subsequent waves, on average, to those individuals observed.

Implementation

All analysis is conducted using R 4.0.2. The data used in this study from the restricted Add Health survey are not made available, but all code and analogous examples are available at https://github.com/ngraetz/multmed_gcomp. Alternatively, the highly accessible *CMAverse* R package (Shi et al. 2021) provides similar g-formula estimators for the CDE and joint mediated effects (PIE/PAI).