On the Causal Interpretation of Race in Regressions Adjusting for Confounding and Mediating Variables

Tyler J. VanderWeele[a] *and Whitney R. Robinson*[b]

Abstract: We consider several possible interpretations of the "effect of race" when regressions are run with race as an exposure variable, controlling also for various confounding and mediating variables. When adjustment is made for socioeconomic status early in a person's life, we discuss under what contexts the regression coefficients for race can be interpreted as corresponding to the extent to which a racial inequality would remain if various socioeconomic distributions early in life across racial groups could be equalized. When adjustment is also made for adult socioeconomic status, we note how the overall racial inequality can be decomposed into the portion that would be eliminated by equalizing adult socioeconomic status across racial groups and the portion of the inequality that would remain even if adult socioeconomic status across racial groups were equalized. We also discuss a stronger interpretation of the effect of race (stronger in terms of assumptions) involving the joint effects of race-associated physical phenotype (eg, skin color), parental physical phenotype, genetic background, and cultural context when such variables are thought to be hypothetically manipulable and if adequate control for confounding were possible. We discuss some of the challenges with such an interpretation. Further discussion is given as to how the use of selected populations in examining racial disparities can additionally complicate the interpretation of the effects.

(*Epidemiology* 2014;25: 473–484)

In observational research to understand health disparities, race/ethnicity is often put in a regression model, and the coefficient estimates are not infrequently interpreted as some measure of health disparity.[1–3] Numerous other sociodemographic, economic, biologic, or psychosocial variables are typically included in these regressions. Some of these variables may be thought of as potentially on the pathway between race/ethnicity and the health outcome. Other variables may be strongly associated with, but seemingly in no sense "caused by," race/ethnicity. The regression coefficient for race/ethnicity is often interpreted as a "health disparity," irrespective of the other variables for which control has been made. However, as we will argue in this article, the interpretation of regression coefficients depends critically on issues of temporal ordering and covariate control.

There have been numerous discussions of approaches to defining the "causal effects of race."[4–9] Some of these focus on specific settings in which "race" itself can be defined as, say, the race perceived on a job application, which can be hypothetically manipulated. In this article, we offer a tentative proposal regarding the general interpretation of a race/ethnicity variable in regression analysis and how this might vary, given the other variables for which control has been made. What we propose certainly does not capture all the subtleties of race/ethnicity in health disparities research, but we hope it can encourage more careful thought in what to include regarding regression models that involve race.

Part of the challenge of interpreting race coefficients causally is that, in the formal causal inference literature, effects are often defined in terms of counterfactual or potential outcomes, which are in turn defined as the outcomes that would result under hypothetical interventions.[10–23] There are, however, no reasonable hypothetical interventions on race when race itself is the exposure. Here, we attempt to provide a causal interpretation of race coefficients in regressions without defining potential outcomes for race itself. When adjustment is made for socioeconomic status early in a person's life, we will see that the race coefficient can sometimes be interpreted as corresponding to the extent to which a racial inequality would remain if various socioeconomic distributions early in life across racial groups could be equalized. When adjustment is also made for adult socioeconomic status, the overall racial inequality can be decomposed into the portion that would be eliminated by equalizing adult socioeconomic status across racial groups and the portion of the inequality that would remain even if adult socioeconomic status across racial groups were equalized. Essentially, we give a plausible causal interpretation of the race coefficient by considering how much a racial inequality could be eliminated by intervening on a different variable, namely socioeconomic status, which may

Submitted 16 May 2013; accepted 17 December 2013.
From the [a]Departments of Epidemiology and Biostatistics, Harvard School of Public Health, Boston, MA; and [b]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC.
T.J.V.W. was supported by National Institutes of Health grant ES017876.
Editors' note: *Related articles appear on pages* 485, 488, *and* 491.
Correspondence: Tyler J. VanderWeele, Departments of Epidemiology and Biostatistics, Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115. E-mail: tvanderw@hsph.harvard.edu.

ISSN: 1044-3983/14/2504-0473
DOI: 10.1097/EDE.0000000000000105

be more manipulable than race. We discuss the possibility of stronger interpretations of race coefficients in regression models and the challenges in doing so.

The elimination of health disparities is one of the US federal government's leading health objectives.[24] Persistently poorer health outcomes for some population groups may indicate violations of US norms of equality of opportunity and individual dignity.[25] Health disparities also limit the economic productivity and well-being of the nation.[25] Understanding the causes of such disparities is central to their being addressed, and we hope that the methodological approach in this article might contribute to that end.

RACE/ETHNICITY: CORRELATES AND COMPONENTS

A racial inequality in a particular health outcome might be said to be present if there is any difference between the outcome for different racial groups. The term "racial disparity" is sometimes used to suggest preventable and unjust racial differences in which a disadvantaged social group experiences worse health than more advantaged groups.[4,5] Here, we use the term "inequality" to indicate any difference, regardless of its modifiability or fairness. Such an inequality may arise because of discrimination; it might also arise because of genetic differences or different cultural contexts. However, to note that there is a difference in a particular outcome is not to explain why the differences are present. To say that there is an inequality, then, is simply to indicate that race and the health outcome are correlated in the population under study.

If we want to discuss the "effects of race," however, we are on shakier ground. In this case, we would want to know that whatever outcome we are studying is in some sense affected by race and not simply affected by some other variable associated with race. The notion of an effect of race is ambiguous: the effects may vary depending on what is meant by race. It may include skin color and its perception by others, parental skin color, and its perception by others, cultural context, or genetic background—all considered separately or jointly.

When the effect of race is under discussion, it will therefore be important to clarify more precisely what aspects of race are intended. Even then, precisely defining and assessing such "race effects" is difficult. Because race is not randomized, whether we consider skin color, parental skin color, genetic background, or cultural context, singly or jointly, all these will likely be correlated with neighborhood income, say, at the time of conception.

In certain studies, we may be able to identify aspects of the effect of race.[6] In family-based studies, particular features of genetic background are effectively randomized, allowing one to estimate the effects of a single genetic variant. In other contexts, if we were interested in assessing race as an indicator of discrimination, we might define the exposure of interest to be the employer's perception of an applicant's race.[7–9] The exposure defined in this manner is subject to conceivable manipulations, such as indicating a particular race on an application. Defining causal effects for an exposure so defined is then relatively unproblematic, and randomized trials can even be conducted to assess this effect and evaluate discrimination.[7–9] However, we cannot, in general hope to conduct a randomized trial that would identify the effect of race as more broadly conceived. If "race/ethnicity" is put in a regression model, this will likely capture the effects of perceived race, along with various other factors such as neighborhood income, quality of schools, and so on, that are correlated with skin color, parental skin color, genetic background, and cultural context.

There has thus been considerable debate as to what, if anything, is meant by the effects of race. The formal causal inference literature has generally conceived of causal effects as a comparison between counterfactuals or potential outcomes.[10,11] Often in the causal inference literature, the position is taken that it is meaningful to speak of a contrast of counterfactual outcomes only to the extent that we can specify an intervention.[12,13] Sometimes this position is associated with the slogan "no causation without manipulation."[14] A literature has begun to develop considering this issue of ill-defined "treatment" or nonmanipulable exposures in more detail.[15–20] However, race is not something we can intervene on, and the associated counterfactual queries generally strike researchers as meaningless. The question of what would a black person's health outcome have been had they been white seems like a strange one to pose. It is sometimes cautioned[21] that one should not discuss the effects of race except in very special circumstances when such effects do correspond to a manipulable variable such as in the examples above of job application studies.

We offer 2 possible interpretations of the effects of race. In the first interpretation, once the components of race are specified, the effect of race corresponds to the joint effects of these specific components for which interventions are at least somewhat more conceivable. There are many challenges with this interpretation, which we discuss below. We refer to this as the "stronger" interpretation of race—stronger in the sense of the assumptions required. In the second (weaker) interpretation, race/ethnicity regression coefficients in a model with certain control variables are interpreted as what would happen to an observed health inequality if certain socioeconomic status distributions were set to something other than what they in fact were. In this weaker interpretation, the intervention will be on a variable that is potentially more manipulable, with the quantity of interest being what such an intervention might do to a health inequality across racial groups.

INTERPRETATION OF RACE/ETHNICITY IN REGRESSION ANALYSIS: CONTROL FOR NONMEDIATING VARIABLES

To simplify discussion further, we assume that only 2 racial groups are under consideration (eg, black and white), although similar remarks could apply to other comparisons. If multiple racial groups were of interest, the methods in this

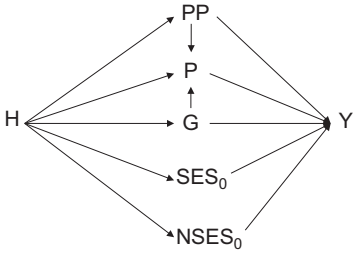

FIGURE 1. Diagram illustrating relations between physical phenotype (*P*), parental physical phenotype (PP), genetic background (*G*), family/parental socioeconomic status (SES_0), neighborhood socioeconomic status ($NSES_0$), history (*H*), and the outcome of interest *Y*.

article could be applied by comparing various racial groups to a single common reference racial group (eg, comparing Asian to white and also comparing black to white).

In trying to understand health inequalities, we might in principle distinguish between forward or "directed pathways" from skin color (or other physical features), parental skin color, or genetic background to the outcome of interest and what we might call "backdoor pathways."[26] The forward or directed pathways from skin color, parental skin color, or genetic background to the outcome are causal pathways from these variables to the outcome, with all edges along the path following the direction of the arrow. The backdoor pathways from skin color, parental skin color, or genetic background to the outcome are pathways that begin with an arrow pointing to skin color, parental skin color, or genetic background.[26] Backdoor pathways might be conceived of as pathways through variables that are associated with skin color, parental skin color, or genetic background (such as family socioeconomic status at the time of conception or birth, neighborhood income, etc). These associations themselves presumably arose from a complex historical process.[22]

Consider the diagram in Figure 1, which is a simplification of a more complex reality but may help illustrate some of the issues concerning interpretation. For now, we assume all variables—physical phenotype (including skin color) (*P*), parental physical phenotype (PP), genetic background (*G*), family/parental socioeconomic status (SES_0), and neighborhood socioeconomic status ($NSES_0$)—are measured at the time of conception. In Figure 1, *H* denotes a complex historical process that gives rise to associations of a person's physical phenotype, parental physical phenotype, and genetic background with the family and neighborhood socioeconomic status into which the person was born. We let *Y* denote the subsequent health outcome of interest. We will also consider below more complicated diagrams that include cultural context. As described below, we will later replace a set of these variables with a self-identified race variable "*R*." We leave "*R*" off of the diagram for now because, before representing it on the diagram, it is important to clarify what is under discussion when the effect of race is being considered.

We use "physical phenotype" as a generic term to include all physical correlates of black versus white race in the United States (such as hair texture) that might be perceived by the person or by others. The effects of physical phenotype include biologic effects of skin color (eg, darker skin protecting against ultraviolet light), the person's understanding of her skin color and other physical features, how this affects her identity and health behaviors, and also, importantly, how others react to the person's physical features (eg, discrimination or feelings of affinity). One objection to language about "an effect of race" or an "effect of physical phenotype" is that such expressions may seem to attribute responsibility for the outcome to the person being discriminated against rather than to the perpetrator of discrimination. While we are sensitive to such linguistic issues, we use expressions such as effects of physical phenotype in the more technical sense associated with causal diagrams.[26] The arrow from physical phenotype to an outcome indicates some causal chain from someone's physical features to the outcome, irrespective of issues of responsibility. It may be the case that an employer discriminates due to an applicant's race in an employment decision; this too is captured in the arrow from physical phenotype to the outcome.

As represented in the diagram, parental physical phenotype may affect the person's subsequent outcome through pathways other than through the individual's own physical phenotype as, for example, might arise if the parents' skin color led to others discriminating against the person as a child. A person's physical phenotype and that of the parent do not, of course, vary independently. Complications can arise with parents of mixed races, adoptions, and albinism, for instance. For simplicity, we will assume that the study population includes only parents of the same race/ethnicity and that physical phenotypes, such as skin color and parental skin color, do in fact coincide. If the groups constituting "mixed race" categories were sufficiently large, then these could themselves be defined as distinct racial groups. The physical phenotype of the parents may of course affect the family (SES_0) and neighborhood socioeconomic status ($NSES_0$) at the time of the child's conception (eg, through discrimination). However, we will denote by the arrow from PP to *Y* the effects of parental physical phenotype on the child's outcome from the time of the child's conception onward. The effects before conception of parental physical phenotype on the outcome (eg, through family and neighborhood SES at the time of conception) will be captured by *H*.

If we put race/ethnicity into a regression model, the interpretation of the coefficient would likely be some combination of the effects of physical phenotype, parental physical phenotype, genetic background, family socioeconomic status, and neighborhood socioeconomic status on the outcome. Suppose, however, that we wanted to isolate the effect of race conceived of as the effects of physical phenotype, parental physical phenotype, and genetic background. The task then

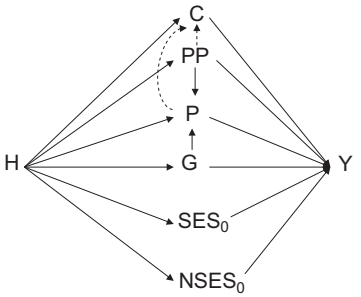

FIGURE 2. Diagram illustrating cultural context (*C*) that may be influenced by physical phenotype (*P*).

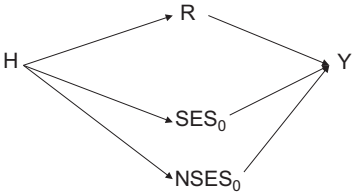

FIGURE 3. Diagram with physical phenotype, parental physical phenotype, genetic background, and cultural context replaced by a race variable (*R*).

would be to control for other variables that were correlated with physical phenotype, parental physical phenotype, genetic background, and the outcome, but not themselves affected by race—ie, we would want to control for variables such as family/parental socioeconomic status and neighborhood socioeconomic status. Essentially, we would want to control for all attributes occurring before conception, but not after conception, because anything occurring after conception could (for instance because of perception of parental physical phenotype) be affected by the variables constituting race. However, to know that we have isolated the forward pathways, we would want to ensure that there were no other variables that both (1) affected the outcome and (2) were correlated with physical phenotype, parental physical phenotype, and genetic background but were not effects of these. We might think of these variables as exposure-outcome confounders with "exposure" here being conceived of as physical phenotype, parental physical phenotype, and genetic background considered jointly. If there were additional variables satisfying conditions (1) and (2), we would want to control for them, as well, to isolate the joint effects of physical phenotype, parental physical phenotype, and genetic background. For example, suppose some aspect of the cultural context (*C*) were correlated with physical phenotype and affected the outcome of interest through pathways independent of SES and neighborhood SES, as in Figure 2. Suppose first that there were no arrow from physical phenotype to cultural context. If we wanted to capture the joint effects of physical phenotype, parental physical phenotype, and genetic background alone, then we would have to control for this cultural context variable as well. If we did not, the regression coefficient for our race/ethnicity variable would also be picking up the effects of culture context associated with physical phenotype.

Of course we may conceive of the effects of race as including those aspects of the cultural context associated with physical phenotype, in which case we would not necessarily want to make regression adjustment for cultural context but allow the race/ethnicity variable to pick this up as well. Cultural context might even be conceived of as being on the pathway from physical phenotype, insofar as physical phenotype may predispose a person toward certain preexisting cultural contexts. If so, we might include an arrow from physical phenotype to cultural context. If this were the case, without adjusting for cultural context, we would be assessing the joint effects of physical phenotype, parental physical phenotype, genetic background, and cultural context. If we did adjust for cultural context, we would have the effects of physical phenotype, parental physical phenotype, and genetic background not through cultural context. In practice, it is unlikely that any measurable variable will adequately capture the cultural context, and thus the race/ethnicity variable will pick up such cultural effects as well.

Once we have decided what is to be included in what we attempt to estimate as the effect of race, we could replace those variables on the diagram with a race variable *R* and leave on the diagram those variables that we would not want to include in the effect of race. For example, from Figure 2, if we wanted to capture in the effect of race the joint effects of physical phenotype, parental physical phenotype, genetic background, and cultural context, we could replace these by our race variable *R* (Figure 3). The diagram then makes clear that to isolate these effects we would need to control for neighborhood and family SES to block the backdoor pathways from our race variable *R* to the outcome. Analytically, we could regress the outcome on our race variable (eg, an indicator for black versus white) and also on neighborhood and family SES. Under the assumption that we have indeed blocked all backdoor pathways by adjusting for neighborhood and family SES, we would obtain with our race coefficient the joint effects (in a sense specified further below) of physical phenotype, parental physical phenotype, genetic background, and cultural context.

If desired, we might not control for neighborhood SES or even family SES in regressions with race as a covariate and thereby allow the race variable to also pick up correlations with and effects of these SES variables and the outcome as well. However, how we interpret the race/ethnicity coefficient will vary according to what is and is not controlled for in the regression model. We could also potentially consider several regression analyses, each with different controls, and each capturing different combinations of the aspects of race.

Formalizing the Interpretation

This still leaves open the question of what precisely is the interpretation of a race/ethnicity coefficient in a regression model with a specific set of control variables in terms

of potential hypothetical interventions. We will consider 2 interpretations of varying strengths. The first interpretation requires stronger assumptions that may often be implausible, and so our focus in the article will be primarily on the second. Suppose that one were willing to conceive of interventions on physical phenotype, parental physical phenotype, genetic background, and cultural context, in a setting such as that of Figure 2, and the health outcome was regressed on race/ethnicity, along with family SES and neighborhood SES. Suppose further that Figure 2 (or Figure 3 with "*R*" indicating "*P*, *PP*, *G*, and *C*") constituted a causal diagram, in that there were no further backdoor pathways from physical phenotype, parental physical phenotype, genetic background, or cultural context through *H* to the outcome *Y* except through variables for which control had been made (eg, family and neighborhood SES). More specifically, suppose that (1) the race variable, *R*, is unassociated with *Y* after controlling for the components of race: physical phenotype, parental physical phenotype, genetic background, cultural context, and family and neighborhood SES and (2) potential associations of physical phenotype, parental physical phenotype, cultural context, and genetic background (even if unmeasured) with the outcome reflect the actual effects of these variables on the outcome once control is made for family and neighborhood SES (see Appendix 1 for greater formality). We argue in Appendix 1 that, under these assumptions, the race coefficient in the regression could be interpreted as the expected difference in health outcomes, for a person with a particular family and neighborhood SES, comparing setting physical phenotype, parental physical phenotype, genetic background, and cultural context to their values from a random draw from the distribution in the white population versus setting these same variables to their values from a random draw from the distribution in the black population. See the article by VanderWeele and Hernán[19] for further discussion of this stronger interpretation of a race coefficient in a regression model. The interpretation may be seen as problematic, in that it may be difficult to conceive of hypothetical interventions on nonmodifiable aspects of physical phenotype, parental physical phenotype, genetic background, and cultural context.

If an investigator objects to the notion of physical phenotype, parental physical phenotype, genetic background, and cultural context being hypothetically manipulable, then a weaker and perhaps more plausible interpretation of an adjusted race coefficient is still possible. It is this weaker interpretation we will focus on. We argue in Appendix 1 that if the coefficients for family and neighborhood SES correspond to the effects of these variables on the outcome, then the coefficient for black race in the regression could be interpreted as the health inequality that would remain between blacks and whites if the family and neighborhood SES distributions (SES_0 and $NSES_0$) of the black population were set equal to that of the white population (eg, by setting SES for each black person to levels randomly chosen from the white SES distribution). Importantly, the coefficient could be interpreted in this way even if one does not want to talk about the effects of race. The coefficient has a causal interpretation without having to define hypothetical interventions on race itself or on any of the variables that might constitute the composite race variable: the coefficient can be interpreted as the resulting health inequality if we were to intervene on family and neighborhood SES. As formalized in Appendix 1, we have a causal interpretation of the race coefficient without defining potential outcomes with respect to race. This is again done by framing the interpretation around interventions on a different variable that may be considered to be more manipulable, namely SES.

Note that the analysis is the same, and thus the estimates will be the same, for the stronger and the weaker interpretations; only the assumptions being made differ. Note, however, that both interpretations require that control be made for family and neighborhood SES. In some contexts, the effects of family and neighborhood SES may be completely confounded by race, in that substantial portions of the SES distributions may not overlap across racial groups (eg, in a particular study in which income disparities were large); if all the lower SES persons were black and all the higher SES persons were white, it would not be possible to distinguish between association due to SES versus race, even if data were available on these variables. This phenomenon is sometimes referred to as "structural confounding,"[23] and it is an issue here as in other analyses examining race and SES. Note also that in practice only certain aspects of family and neighborhood SES will be used in any given analysis, and so the effects here would have to be interpreted as the resulting health inequality of setting the distributions of the particular SES variables used in the analysis equal across racial groups.

INTERPRETATION OF RACE/ETHNICITY IN REGRESSION ANALYSIS: CONTROL FOR MEDIATING VARIABLES

In health disparities research, it is not infrequent to control for socioeconomic status (either individual or neighborhood-level) later in life, in addition to or instead of socioeconomic status at birth. Unlike factors described above, these factors temporally occur after race. These factors might then be mediators of the effect of race, ie, variables on forward pathways from race to the outcome. Controlling for mediating factors changes the interpretation of regression coefficients and purported effect estimates. The interpretation of the role of SES later in life is arguably distinct from the interpretation of SES in childhood or at birth. Again, socioeconomic status later in life is arguably on the pathway from physical phenotype, parental physical phenotype, and genetic background, not simply correlated with them due to some prior historical process as is the case for family or neighborhood socioeconomic status measured at conception. If the aim of an analysis were to assess the effects of race conceived of as the overall joint effects of physical phenotype, parental physical

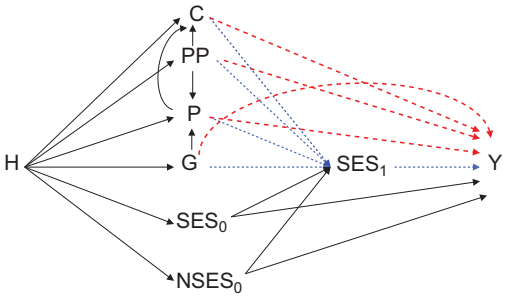

FIGURE 4. Diagram with adult socioeconomic status (SES_1) and the pathways from race components to the outcome (*Y*) through adult SES (the blue dotted pathways) and not through SES (the red dashed pathways).

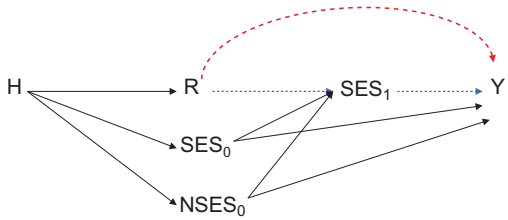

FIGURE 5. Effects of race through adult SES (the blue dotted pathways) and not through SES (the red dashed pathways), with physical phenotype, parental physical phenotype, genetic background, and cultural context replaced by a race variable (*R*).

phenotype, genetic background, and cultural context, then one would not want to adjust for socioeconomic status later in life. Some of the effect would potentially be blocked if control were made for such a variable measured later in life.

Alternatively, control for SES later in life is perhaps sometimes done to assess the extent to which health inequalities across racial groups are explained by differing SES levels later in life. Consider the diagram in Figure 4 where SES_1 indicates individual SES in early adulthood, at age 25 years say. Suppose we were once again interested in the joint effects of physical phenotype, parental physical phenotype, genetic background, and cultural context, but that now we also wanted to distinguish the extent to which these joint effects were mediated by individual SES in early adulthood (the blue dotted paths) and the extent to which they were through other pathways (the red dashed paths). If we wanted to capture the effects of race conceived of as the joint effects of physical phenotype, parental physical phenotype, genetic background, and cultural context, we could once again replace these with a single variable R on the diagram as in Figure 5. As argued above, under the stronger interpretation, the coefficient for race/ethnicity in a regression of the outcome of interest might be interpreted as an overall effect of physical phenotype, parental physical phenotype, genetic background, and cultural context if we were able to control for family and neighborhood SES early in life (and other variables that may lie on backdoor pathways). This overall effect would thus give us the blue and red pathways combined. To separate these pathways, one would want to estimate the "direct effects" of physical phenotype, parental physical phenotype, genetic background, and cultural context not through adult SES and the effects of these variables "mediated by" adult SES.

There is now a body of work in the causal inference literature[26–36] on estimating direct and indirect effects. In the context of well-defined manipulable exposures and mediators, estimating such effects requires baseline control for exposure-outcome, mediator-outcome, and exposure-mediator confounders.[26] Uncontrolled confounders of the mediator-outcome relationship can lead to substantial biases in these effects.[26,27,31] The application of the mediation analysis literature to the health disparities context is potentially problematic if the effects of race are thought to be not well-defined.[22]

As before, we could potentially proceed in 1 of 2 ways with regard to interpretation. Under a stronger interpretation in which the effects of race were conceived of as the joint effects of physical phenotype, parental physical phenotype, genetic background, and cultural context, the ideas from the causal inference literature concerning direct and indirect effects could be applied. However, this would again require counterfactuals that set physical phenotype, parental physical phenotype, genetic background, and cultural context to specific values, which may not be thought to be plausible and therefore will not be pursued further here. An alternative weaker and perhaps more plausible interpretation within the context of health disparities research, however, arises from hypothetical interventions on the SES distributions themselves, which we will now describe.

Suppose that the methods from the causal inference literature for direct and indirect effects are used in the health disparities context with race as the exposure, adult socioeconomic status as the mediator, and some adult outcome, with individual and neighborhood socioeconomic status at birth as additional covariates. Suppose that we have controlled for sufficient variables such that the association between adult SES and the outcome actually reflects the effects of adult SES on the outcome. This is essentially an analog of the mediator-outcome confounding control assumption in the literature on direct and indirect effects (no analog of the other assumptions is necessary here because we are not intervening on the exposure, as discussed in Appendix 1). We argue in Appendix 1 that if these assumptions hold, the direct effect obtained for race not through adult SES (when also controlling for family SES and neighborhood SES at conception or early in life) could be interpreted as the health inequality that would remain for persons with a particular early family and neighborhood SES level, if within this population, the adult SES distribution of the black population were set equal to that of the white population (eg, by setting SES for each black individual to levels randomly chosen from the white SES distribution). We might refer to this as a "direct-effect racial inequality measure" not through adult

SES (ie, how much of the inequality remains after accounting for adult SES). We also argue that what is estimated as an indirect or mediated effect can be interpreted as how the health outcomes for the black population with a particular early family and neighborhood SES level would change if the adult SES distribution of this black population were set equal to that of the black population versus that of the white population. We might refer to this as a "mediated racial inequality measure" through adult SES (ie, how much of the inequality is due to difference in adult SES). Moreover, we show that the overall health inequality for those with a particular early family and neighborhood SES level is equal to the sum of these "direct" and "mediated" racial inequality measures. We again can interpret coefficients in this way without having to define potential outcomes with respect to race or without defining what might be meant by the effects of race. This is once again done by framing the interpretation around interventions on a different variable that may manipulable, adult SES.

Importantly, however, these direct and mediated racial inequality measures will always be with respect to the particular adult SES measurement used in the analysis. Socioeconomic status has numerous dimensions, and no measurement will adequately capture all of these.[37] Even under the assumptions above, the mediated racial inequality measure will capture only the portion of the racial inequality due to the particular measure of adult SES used in the analysis, not adult SES in its entirety. The interpretation of the effects again corresponds to equalizing across racial groups the distributions of the SES variable or variables used in the analysis.

A number of methods have been proposed to estimate these direct and indirect effects.[28–36] However, sometimes the approach of simply including the "mediator variable" (here adult SES) in the model will suffice. In particular, if the outcome is continuous and there is no statistical interaction between the exposure variable (race) and the mediator variable (adult SES), then the coefficient for race in the model that includes adult SES (and the control variables) will correspond to a direct effect, and the difference in the coefficients for race in the models without versus with adult SES will correspond to the mediated effect.[29] For a binary outcome with logistic regression, provided that the outcome is rare (or if a log-linear model is used with a common outcome), and if there is no statistical interaction between race and adult SES, then once again the coefficient for race in the model that includes adult SES (and the control variables) when exponentiated will correspond to a direct effect odds ratio, and the difference in the coefficients for race in the models without versus with adult SES when exponentiated will correspond to the mediated effect odds ratio.[30] On the odds ratio scale for logistic regression, the overall racial inequality measure will decompose into a product (rather than the sum) of the direct and mediated racial inequality measures. As noted above, the interpretation of the direct and indirect effect measures will hold if covariate control suffices for the associations between adult SES and the outcome to actually reflect the effects of adult SES on the outcome; again this is the analog of the mediator-outcome confounding control assumption in the causal inference literature on direct and indirect effects.

The methods for direct and indirect effects[28–33] can, however, also be used to obtain direct and mediated effect estimates even when there is potential interaction between race and adult SES (eg, if the effects of adult SES differ by racial group). And indeed there is some theoretical and empirical evidence for such interaction between race and SES for at least some health outcomes.[38,39] When using newer approaches to obtain direct and mediated effect estimates even in the presence of interaction, the interpretation of these effect estimates would again be that given above. The methods for direct and indirect effects from the causal inference literature can also allow for interactions between race and other variables,[32] such as sex or year of birth, if these are thought to be present.

ILLUSTRATION

We provide a simple illustration, not intended to be a full rigorous analysis, of the approaches described above with an example concerning black-white differences in body mass index (BMI) among US women. Data come from the National Longitudinal Study of Adolescent Health (Add Health), a nationally representative cluster-sample survey of US public and private school students enrolled in grades 7 through 12.[40] At the baseline survey, detailed questionnaires were administered to each student and to the student's primary cohabitating caregiver (preferentially a cohabitating woman). We analyzed data from non-Hispanic white and black women who completed the 2008 follow-up visit. Respondents were 24 to 32 years of age. Race and ethnicity were self-reported. Respondents' heights and weights were measured by trained interviewers and used to calculate the outcome, BMI (kg/m^2).[41]

Childhood family SES was defined by continuous maternal education, self-reported by the respondent's biological or adoptive mother when the respondent was in secondary school. Childhood neighborhood SES was defined from the US census as the percentage of college graduates among the adults 25 years of age or older in the census block in which the respondent lived at the baseline survey. Adult SES was defined by years of attained education in 2008 (range: 6–21). All analyses controlled for age. Models for BMI also were fit controlling for (1) measures of childhood family SES and childhood neighborhood SES, (2) adult SES, and (3) the interaction of race and childhood family SES. All models were weighted to account for Add Health's complex survey sampling and nonresponse.[40]

The overall excess of BMI in black versus white women was 3.74 BMI units (95% confidence interval [CI] = 2.90 to 4.58). When control was made for childhood SES (measured by years of maternal education), this difference became 3.54 (2.41 to 4.36). When adjustment was further made for early neighborhood SES (measured by percentage of adults

with college degrees), this became 3.20 (1.65 to 3.99). Under a stronger interpretation, this difference of 3.20 BMI units could be interpreted as the effects of physical phenotype and parental physical phenotype, genetic background, and cultural context if we thought we had adequately adjusted for confounding for the effects of these variables; in this illustration, this seems unlikely, given that our family and neighborhood SES measures capture only part of the desired underlying construct. Under the weaker interpretation, the estimate of 3.20 corresponds to the racial inequality had we set the distributions of our early family and neighborhood SES distribution in black women to be what they were among white women.

When adjustment is also made for adult SES (measured by adult education), the difference is attenuated only slightly to 3.17 (2.38 to 3.96). Here, ignoring potential interaction between race and adult SES, the "direct-effect" racial inequality measure is 3.17 and the "mediated effect" racial inequality measure (through adult education) is only 0.03 (95% CI = −0.08 to 0.14). From these data, it appears that only about 1% of the BMI difference would be eliminated if adult SES distributions were equalized; thus, most of the racial inequality does not seem to be due to differences in our measure of adult SES, namely years of education. When allowing for interaction between race and adult education, the estimates remain virtually unchanged. Although some of the initial racial inequality is explained by these measures of neighborhood and family SES in childhood, very little of it is explained or mediated by years of education attained in adulthood.

DISCUSSION

We have considered the causal interpretations of the race coefficient in regression models controlling for confounding and mediating variables, and we have provided interpretations of those coefficients that do not require defining potential outcomes on race itself. The interpretation provided is as a racial inequality that would remain if various socioeconomic status distributions across racial groups were equalized. This interpretation is retained without requiring hypothetical manipulation on race or its components (eg, physical phenotype, parental physical phenotype, genetic background, and cultural context). This interpretation was accomplished by framing the interpretation around interventions on various SES distributions, which may be more manipulable. We discussed also a stronger interpretation of the race coefficient when interventions on various components of race, eg, physical phenotype, parental physical phenotype, genetic background, and cultural context, was thought possible, but we noted that such interventions may be more difficult to conceive.

Our discussion has focused on differences in outcomes across racial groups. Sometimes such differences are examined for selected populations, such as racial inequalities for pregnant women or racial inequalities in outcomes for those with asthma. Such selected populations create further challenges for the interpretation of race coefficients in regression models and are discussed further in more detail in Appendix 2.

A similar approach might also be used with other nonmanipulable exposure such as sex. The approach might also be used with factors other than socioeconomic status that may differ across racial groups.

Importantly, we have shown that the interpretation of the race coefficient differs depending on whether variables such as individual and neighborhood socioeconomic status are controlled for at birth or later in life. Comparisons could be made across a variety of socioeconomic variables or other factors to attempt to determine what interventions either early or later in life might most substantially eliminate later health inequalities. However, when these various SES variables are themselves strongly correlated with one another, and when control is not made for the others, it may be difficult to isolate effects. To interpret the effects as we have in this article, the SES variables themselves, as we have noted, must be unconfounded. An investigator need not restrict attention to one analysis but may run a series of regression analyses or use modern methods for direct and indirect effect, accounting also for interaction between race and socioeconomic status, to gain insight into the sources of disparities.

ACKNOWLEDGMENTS

We thank the editors and 3 anonymous reviewers for helpful comments.

REFERENCES

1. Wu YW, Xing G, Fuentes-Afflick E, Danielson B, Smith LH, WM. Racial, ethnic, and socioeconomic disparities in the prevalence of cerebral palsy. *Pediatrics*. 2011;127:e674–e681.
2. Foster EM. Medicaid and racial disparities in health: the issue of causality. A commentary on Rose et al. *Soc Sci Med*. 2010;70:1271–1273; discussion 1274.
3. Naimi AI, Kaufman JS, Howe CJ, Robinson WR. Mediation considerations: serum potassium and the racial disparity in diabetes risk. *Am J Clin Nutr*. 2011;94:614–616.
4. Braveman P. Health disparities and health equity: concepts and measurement. *Annu Rev Public Health*. 2006;27:167–194.
5. Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care. Unequal treatment: confronting racial and ethnic disparities in health care. In: Smedley BD, Stith AY, Nelson AR, eds. *Board on Health Sciences Policy*. Washington, DC: Institute of Medicine of the National Academies; 2003:1–243.
6. Kaufman JS, Cooper RS. Commentary: considerations for use of racial/ethnic classification in etiologic research. *Am J Epidemiol*. 2001;154:291–298.
7. Bertrand M, Mullainathan S. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am Econ Rev.* 2004;94:991–1013.
8. Butler DM, Broockman DE. Do politicians racially discriminate against constituents? A field experiment on state legislators. *Am J Pol Sci*. 2011;55:463–477.
9. Sen M, Wasow O. How and when to make causal claims based on race or ethnicity. *Technical Report*. 2013.
10. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educational Psychol.* 1974;66:688–701.
11. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Ann Stat.* 1978;6:34–58.
12. Robins JM, Greenland S. Comment on "Causal inference without counterfactuals." *JASA.* 2000;95:477–482.

13. Hernán MA. Hypothetical interventions to define causal effects: afterthought or prerequisite? *Am J Epidemiol*. 2005;162:618–620.
14. Holland P. Statistics and causal inference. *JASA*. 1986;81:945–960.
15. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009;20:3–5.
16. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20:880–883.
17. Pearl J. On the consistency rule in causal inference: axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21:872–875.
18. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology*. 2011;22:368–377.
19. VanderWeele TJ, Hernán MA. Causal effects and natural laws: towards a conceptualization of causal counterfactuals for non-manipulable exposures with application to the effects of race and sex. In: Berzuini C, Dawid P, Bernardinelli L, eds. *Causal Inference: Statistical Perspectives and Application*. West Sussex, UK: Wiley and Sons; 2012:101–113.
20. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. *J Causal Inference.* 2013;1:1–20.
21. Greiner D, Rubin DB. Causal effects of perceived immutable characteristics. *Rev Econ Stat*. 2011;93:775–785.
22. Kaufman JS. Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. *Soc Sci Med*. 2008;66:1659–1669.
23. Messer LC, Oakes JM, Mason S. Effects of socioeconomic and racial residential segregation on preterm birth: a cautionary tale of structural confounding. *Am J Epidemiol*. 2010;171:664–673.
24. U.S. Department of Health and Human Services. *Office of Disease Prevention and Health Promotion. Healthy People*. 2020. Washington, DC. Available at: http://www.healthypeople.gov/2020/about/disparities-About.aspx. Accessed 26 September 2013.
25. Williams DR, Jackson PB. Social sources of racial disparities in health. *Health Aff*. 2005;24:325–334.
26. Pearl J. Direct and indirect effects. In: *Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence*. San Francisco, Calif: Morgan Kaufmann; 2001:411–420.
27. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*. 1992;3:143–155.
28. van der Laan MJ, Petersen ML. Direct effect models. *Int J Biostat.* 2008;4:Article 23.
29. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface*. 2009;2:457–468.
30. VanderWeele TJ, Vansteelandt S. Odds ratios for mediation analysis with a dichotomous outcome. *Am J Epidemiol.* 2010;172:1339–1348.
31. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology*. 2010;21:540–551.
32. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods*. 2010;15:309–334.
33. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychol Methods*. 2013;18:137–150.
34. Lange T, Vansteelandt S, Bekaert M. A simple unified approach for estimating natural direct and indirect effects. *Am J Epidemiol*. 2012;176:190–195.
35. Vansteelandt S, Bekaert M, Lange T. Imputation strategies for the estimation of natural direct and indirect effects. *Epidemiol Method*. 2012;1:131–158.
36. Tchetgen Tchetgen EJ, Shpitser I. Semiparametric theory for causal mediation analysis: efficiency bounds, multiple robustness, and sensitivity analysis. *Ann Stat*. 2012;40:1816–1845.
37. Kaufman JS, Cooper RS, McGee DL. Socioeconomic status and health in blacks and whites: the problem of residual confounding and the resiliency of race. *Epidemiology*. 1997;8:621–628.
38. Sánchez-Vaznaugh EV, Kawachi I, Subramanian SV, Sánchez BN, Acevedo-Garcia D. Do socioeconomic gradients in body mass index vary by race/ethnicity, gender, and birthplace? *Am J Epidemiol*. 2009;169:1102–1112.
39. Chang VW, Lauderdale DS. Income disparities in body mass index and obesity in the United States, 1971-2002. *Arch Intern Med*. 2005;165:2122–2128.
40. Harris KM, Halpern CT, Whitsel E, et al. The National Longitudinal Study of Adolescent Health: Research Design. 2009. Available at: http://www.cpc.unc.edu/projects/addhealth/design. Accessed 4 April 2013.
41. World Health Organization. Physical status: the use and interpretation of anthropometry. Report of a WHO Expert Committee. *World Health Organ Tech Rep Ser*. 1995;854:1–452.
42. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615–625.
43. VanderWeele TJ, Robins JM. Minimal sufficient causation and directed acyclic graphs. *Ann Stat*. 2009;37:1437–1465.
44. Evans AT, Sadowski LS, VanderWeele TJ, et al; CHIRAH Study Group. Ethnic disparities in asthma morbidity in Chicago. *J Asthma*. 2009;46:448–454.

APPENDIX 1: PROOFS

Interpretation of Total Effects

Let R denote the race/ethnicity variable used in the regression. Let $R = 1$ indicate black and $R = 0$ indicate white. Let A = (P, PP, G, C) denote the collection of physical phenotype, parental physical phenotype, genetic background variables, and cultural context variables. Let Y denote the health outcome. Let $X = (\mathrm{SES}_0, \mathrm{NSES}_0)$ denote family and neighborhood SES at the time of conception or early in life (or more generally variables thought to be associated with A and Y but not affected by A). Suppose we were to fit the following regression:

$$\mathrm{E}[\mathrm{Y}|\mathrm{r}, \mathrm{x}] = \beta_0 + \beta_1 \mathrm{r} + \beta_2' \mathrm{x}$$

For the weaker interpretation, let $G(0)$ denote a random draw of early family and neighborhood SES (ie, the variables X) of the white population. Let Y_x denote an individual's counterfactual outcome if their early family and neighborhood SES were set to x. Then $E[Y_{G(0)}|R = 1]$ would denote the expected outcome in the black population if for each individual their early family and neighborhood SES were set to a value from a random draw from their distribution in the white population. Note that $\mathrm{pr}(G(0) = x) = \mathrm{pr}(x|R = 0)$ and also because $G(0)$ is random, $\mathrm{pr}(G(0) = x) = \mathrm{pr}(G(0) = x|R = 1)$. If the effects of family and neighborhood SES on the outcome are unconfounded conditional on R, ie, $E[Y_x|R = 1] = E[Y|R = 1, x]$, so that the associations of family and neighborhood SES with the outcome correspond to the effects of these variables on the outcome then, from the regression model, we have that:

$$\beta_1 = \mathrm{E}[\mathrm{Y}|\mathrm{R} = 1, \mathrm{x}] - \mathrm{E}[\mathrm{Y}|\mathrm{R} = 0, \mathrm{x}]$$

If we sum this over the distribution of $\mathrm{pr}(x|R = 0)$, we get

$$\beta_1 = \textstyle\sum_{\mathrm{x}} \mathrm{E}[\mathrm{Y}|\mathrm{R} = 1, \mathrm{x}]\, \mathrm{pr}(\mathrm{x}|\mathrm{R} = 0) - \mathrm{E}[\mathrm{Y}|\mathrm{R} = 0, \mathrm{x}]\, \mathrm{pr}(\mathrm{x}|\mathrm{R} = 0)$$

$$\beta_1 = \textstyle\sum_{\mathrm{x}} \mathrm{E}[\mathrm{Y}|\mathrm{R} = 1, \mathrm{x}]\, \mathrm{pr}(\mathrm{x}|\mathrm{R} = 0) - \mathrm{E}[\mathrm{Y}|\mathrm{R} = 0]$$

$$\beta_1 = \textstyle\sum_{\mathrm{x}} \mathrm{E}[\mathrm{Y}_{\mathrm{x}}|\mathrm{R} = 1]\, \mathrm{pr}(\mathrm{G}(0) = \mathrm{x}) - \mathrm{E}[\mathrm{Y}|\mathrm{R} = 0]$$

$$\beta_1 = \textstyle\sum_{\mathrm{x}} \mathrm{E}[\mathrm{Y}_{\mathrm{x}}|\mathrm{R} = 1, \mathrm{G}(0) = \mathrm{x}]\, \mathrm{pr}(\mathrm{G}(0) = \mathrm{x}|\mathrm{R} = 1) - \mathrm{E}[\mathrm{Y}|\mathrm{R} = 0]$$

$$\beta_1 = E\left[Y_{G(0)}|R = 1\right] - E[Y|R = 0].$$

Thus, the race coefficient in the regression could be interpreted as the racial inequality that would remain if the family and neighborhood SES distribution of the black population were set equal to that of the white population. Note that under this weaker interpretation, we have defined potential outcomes for Y based on interventions on early family and neighborhood SES but not on race.

For the stronger interpretation, let Y_a be the outcome that would have been observed for an individual if physical phenotype, parental physical phenotype, genetic background, and cultural context were set to a. We then have:

$$\beta_1 = E[Y|R = 1, x] - E[Y|R = 0, x]$$

$$= \sum_a E[Y|R = 1, a, x]\, pr(a|R = 1, x) - \sum_a E[Y|R = 0, a, x]\, pr(a|R = 0, x).$$

If R is independent of Y conditional on A and X, then we have that this equals:

$$= \sum_a E[Y|a, x]\, pr(a|R = 1, x) - \sum_a E[Y|a, x]\, pr(a|R = 0, x)$$

If the effects of A on Y are unconfounded conditional on x, ie, if $E[Y|a, x] = E[Y_a|x]$, so that the associations between A and Y conditional on X reflect the effects of A then this equals:

$$= \sum_a E\left[Y_a|x\right] pr(a|R = 1, x) - \sum_a E\left[Y_a|x\right] pr(a|R = 0, x)$$

Thus, the race coefficient in the regression could be interpreted as the expected difference in health outcomes, for those with early family and neighborhood SES level of x, between setting physical phenotype, parental physical phenotype, genetic background, and cultural context to their values from a random draw from an individual in the white population versus settings these same variables to their values from a random draw from an individual in the black population.

Interpretation of Direct and Mediated Effects

Let R denote the race/ethnicity variable used in the regression. Let $R = 1$ indicate black and $R = 0$ indicate white. Let $A = (P, PP, G, C)$ denote the collection of physical phenotype, parental physical phenotype, genetic background, and cultural context variables. Let M denote adult SES. Let Y denote the health outcome. Let $X = (SES_0, NSES_0)$ denote family and neighborhood SES at the time of conception or early in life. Let $H_x(0)$ be a random draw from the adult SES distribution of the white population with baseline covariates x. Let Y_m denote an individual's random counterfactual outcome if his or her adult SES were set to m. Then, $E\left[Y_{H_x(0)}|R = 1, x\right]$ denotes the expected outcome for a black individual with early family and neighborhood SES of x if their adult SES were set to a random draw from that of the white population with early family and neighborhood SES of x. Note that $pr(H_x(0) = m|x, r) = pr(H_x(0) = m) = pr(m|R = 0, x)$. Suppose also that the effects of M on Y are unconfounded conditional on (R, X), ie, $E[Y_m|R = 1, x] = E[Y|R = 1, m, x]$, so that the associations between adult SES and the outcome reflect the actual effects of adults SES. Methods from the mediation analysis literature for the natural direct effect[26,28,30,32] conditional on X with R as the exposure, M as the mediator, and Y as the outcome effectively estimate:

$$\sum_m E[Y|R = 1, m, x]\, pr(m|R = 0, x) - \sum_m E[Y|R = 0, m, x]\, pr(m|R = 0, x)$$

$$= \sum_m E\left[Y_m|R = 1, H_x(0) = m, x\right] pr\left(H_x(0) = m|R = 1, x\right) - E[Y|R = 0, x]$$

$$= E\left[Y_{H_x(0)}|R = 1, H_x(0), x\right] - E[Y|R = 0, x]$$

Thus, the direct effect that is obtained for race not through adult SES (when also controlling for family SES and neighborhood SES at conception or early in life) could be interpreted as the racial inequality that would remain for individuals with early family and neighborhood SES level of x, if within this population, the adult SES distribution of the black population were set equal to that of the white population.

Methods from the mediation analysis literature for the natural indirect effect[26,28,30,32] conditional on X with R as the exposure, M as the mediator, and Y as the outcome effectively estimate:

$$\sum_m E[Y|R = 1, m, x] pr(m|R = 1, x) - \sum_m E[Y|R = 1, m, x] pr(m|R = 0, x)$$

Similarly, as above, let $H_x(1)$ be a random draw from the adult SES distribution of the black population with baseline covariates x so that $E[Y_{Hx(1)}|R = 1, x]$ denotes the expected outcome for a black individual with early family and neighborhood SES of x if their adult SES were set to a random draw from that of the black population with early family and neighborhood SES of x. Note that $pr(H_x(1) = m) = pr(H_x(1) = m|x, r) = pr(m|R = 1, x)$. If the effects of M on Y are unconfounded conditional on (R, X), ie, $E[Y_m|R = 1, x] = E[Y|R = 1, m, x]$, so that the associations between adult SES and the outcome reflect the actual effects of adult SES, then we have:

$$\sum_m E[Y|R = 1, m, x] pr(m|R = 1, x) - \sum_m E[Y|R = 1, m, x] pr(m|R = 0, x)$$

$$= \sum_m E\left[Y_m|R = 1, H_x(1) = m, x\right] pr\left(H_x(1) = m|R = 1, x\right) - \sum_m E\left[Y_m|R = 1, H_x(0) = m, x\right] pr\left(H_x(0) = m|R = 1, x\right)$$

$$= E\left[Y_{H_x(1)}|R = 1, x\right] - E\left[Y_{H_x(0)}|R = 1, x\right].$$

The mediated effect can thus be interpreted as how the health outcomes for the black population with early family and neighborhood SES of x would change if the adult SES distribution of this black population were set equal to that of the black population versus that of the white population.

The overall racial inequality measure for those with early family and neighborhood SES of x is given by:

$$\begin{aligned}
&\mathrm{E}[\mathrm{Y}|\mathrm{R}=1,\mathrm{x}]-\mathrm{E}[\mathrm{Y}|\mathrm{R}=0,\mathrm{x}]\\
&=\textstyle\sum_{\mathrm{m}}\mathrm{E}[\mathrm{Y}|\mathrm{R}=1,\mathrm{m},\mathrm{x}]\,\mathrm{pr}(\mathrm{m}|\mathrm{R}=1,\mathrm{x})\\
&\quad-\textstyle\sum_{\mathrm{m}}\mathrm{E}[\mathrm{Y}|\mathrm{R}=0,\mathrm{m},\mathrm{x}]\,\mathrm{pr}(\mathrm{m}|\mathrm{R}=0,\mathrm{x})\\
&=\textstyle\sum_{\mathrm{m}}\mathrm{E}[\mathrm{Y}|\mathrm{R}=1,\mathrm{m},\mathrm{x}]\mathrm{pr}(\mathrm{m}|\mathrm{R}=1,\mathrm{x})\\
&\quad-\textstyle\sum_{\mathrm{m}}\mathrm{E}[\mathrm{Y}|\mathrm{R}=1,\mathrm{m},\mathrm{x}]\mathrm{pr}(\mathrm{m}|\mathrm{R}=0,\mathrm{x})\\
&\quad+\textstyle\sum_{\mathrm{m}}\mathrm{E}[\mathrm{Y}|\mathrm{R}=1,\mathrm{m},\mathrm{x}]\,\mathrm{pr}(\mathrm{m}|\mathrm{R}=0,\mathrm{x})\\
&\quad-\textstyle\sum_{\mathrm{m}}\mathrm{E}[\mathrm{Y}|\mathrm{R}=0,\mathrm{m},\mathrm{x}]\mathrm{pr}(\mathrm{m}|\mathrm{R}=0,\mathrm{x})\\
&=\left\{\mathrm{E}\left[\mathrm{Y}_{\mathrm{H_x}(1)}|\mathrm{R}=1,\mathrm{x}\right]-\mathrm{E}\left[\mathrm{Y}_{\mathrm{H_x}(0)}|\mathrm{R}=1,\mathrm{x}\right]\right\}\\
&\quad+\mathrm{E}\left[\mathrm{Y}_{\mathrm{H_x}(0)}|\mathrm{R}=1,\mathrm{x}\right]-\mathrm{E}[\mathrm{Y}|\mathrm{R}=0,\mathrm{m},\mathrm{x}].
\end{aligned}$$

where the second equality is obtained by adding and subtracting $\sum_m E[Y|R = 1, m, x]$ pr $(m|R = 0, x)$ and, in the third equality, the 2 expressions are simply the direct effect and mediated effect disparities measures given above. Note that although the empirical expressions here are the same as those that are used for so-called natural direct and indirect effects,[26,27] the assumptions required here for identification are much weaker than those for natural direct and indirect effects because the "mediator" is not being fixed to the level it would have had for that individual under a counterfactual scenario, as it is for natural direct and indirect effects, but it is rather being fixed randomly to a value from an observed distribution, namely that of the other racial group. Note that we can define these effects and have this decomposition without defining potential outcomes for Y with regard to race; we instead defined, as above, potential outcomes for Y based on interventions on adult SES.

A similar interpretation would hold for binary outcomes on an odds ratio scale provided the outcome is rare.[28] If the outcome is continuous and there are no statistical interactions between R and M, then the coefficient for R in the model that includes M (and X) will give the empirical quantity used to estimate the direct effect, and the difference in the coefficients for race in the models without versus with adult SES will give the empirical quantity used to estimate the mediated effect.[29] For a binary outcome with logistic regression, provided that the outcome is rare (or if a log-linear model is used with a common outcome), and if there are no statistical interactions between R and M, then once again the coefficient for R in the model that includes M (and X) will give the empirical quantity used to estimate the direct effect, and the difference in the coefficients for race in the models without versus with adult SES will give the empirical quantity used to estimate the mediated effect.[30]

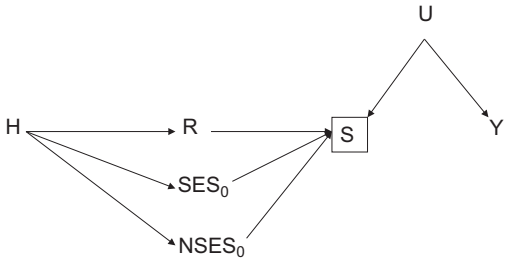

FIGURE 6. Diagram illustrating bias in selected populations (S) in associations between race (R) and outcome (Y) that can result because of common causes of the variable defining the population (S) and the outcome (Y).

APPENDIX 2: SELECTED POPULATIONS

Our discussion thus far has considered "unselected" populations; that is to say, cohorts of different racial groups followed up to compare differences in some health outcome. It is not infrequent, however, to also consider health disparities among selected populations. For example, racial inequalities might be examined for birth outcomes for pregnant women, for survival following the onset of breast cancer, or for severe asthma exacerbation among children with asthma. Here, the populations of interest are defined by some variable, event, or shared characteristic (eg, pregnancy, breast cancer, or asthma). So long as the exposure of interest occurs after the event or characteristic defining the population, the analysis of such selected populations is unproblematic. However, if the exposure of interest occurs before the event or variable defining the population, this can then bias comparisons across exposure groups if the exposure itself affects the variable/event defining the population.

In the context of health disparities research, if race constitutes the exposure variable and if race (eg, physical phenotype, parental physical phenotype, genetic background, cultural context) also affects the likelihood of pregnancy, breast cancer, or asthma, then comparisons of outcomes across racial groups within the selected population may give associations that arise from working with a selected population rather than because physical phenotype, parental physical phenotype, genetic background, or cultural context have effects on the outcome. To see this, consider the relations in Figure 6. As before, suppose we wish to assess the effects of race conceived of as the joint effects of physical phenotype, parental physical phenotype, genetic background, and cultural context (denoted by our race variable R, with control for neighborhood and family SES to isolate these effects). Let S denote the variable defining the population (eg, pregnancy). The box around S indicates that we are conditioning on the event being present ($S = 1$). In Figure 6, race does not

affect the outcome *Y* (eg, none of physical phenotype, parental physical phenotype, genetic background, cultural context—or even neighborhood and family SES—affect the outcome). However, race does affect the likelihood of the event defining the population *S*. Suppose also that there were a common cause *U* of *S* and *Y*; for example, if *S* indicated pregnancy and *Y* were acne, *U* might be age. If we were to look at associations between *R* and *Y* conditional on *S*, we would find associations even though there were no effects of *R* on *Y*.

This is because we are conditioning on a variable that is a common effect *S* of (1) the exposure variable *R* and also (2) a variable associated with *Y*, namely *U*.[42] Doing so introduces spurious correlation, sometimes known as collider stratification bias. Here, if analysis were restricted to pregnant women, then even if race did not affect acne, it might look like, among pregnant women, race affected acne, but this would be black because black women are pregnant at younger ages and those who are younger have more acne. As discussed further below if control could be made for the common cause(s) *U* of the outcome *Y* and the variable *S* defining the population, then such biases would be eliminated. However, without such control, in cases in which *R* itself does in fact also affect *Y*, such bias will distort associations between *R* and *Y* once we condition on the event *S* being present. This renders any of the interpretations for the coefficients of race in regression models problematic.

Although giving a causal interpretation to regression coefficients involving race was difficult even in unselected population, the issues of interpretation become even more difficult in selected populations. Several responses and approaches to address such issues in selected populations are, however, possible. First, if what we are interested in is only description, then it may still be of interest that there are racial differences in a health outcome even if these do not necessarily correspond to something that can be interpreted causally. For example, we may be interested in whether pregnancy outcomes vary for black versus white mothers, even if these associations may be due to different characteristics of white and black women who become pregnant rather than to the effects of race (eg, discrimination in response to physical phenotype or differences in genetic background) on birth outcomes.

Second, if we do want to causally interpret associations between race and a health outcome in a selected population, we could still do so if either (1) race did not affect the likelihood of the event defining the population, ie, no arrows from *R* (or its components in Figure 5) to *S* or (2) if we were able to control for common causes (eg, *U* in the diagram) of the event *S* defining the population and the outcome *Y* or if there were no such common causes. In these cases, we could maintain the causal interpretations of the associations between race and the health outcome given above. Third, we could shift focus and look at racial differences in outcomes across the entire population rather than in a selected population; for example, we could look at acne differences for all women not simply pregnant women.

Finally, there may be other methodological approaches that can help in these settings of selected populations. In some cases, we may be able to reason about the direction of the bias that results from collider stratification. For example, if both *R* and *U* affect *S* in the same direction, we might expect *R* and *U* to be negatively correlated conditional on *S* (eg, if in some cases *S* is present when either *R* or *U* is, then if $R = 0$ and $S = 1$ we would know $U = 1$ and vice versa). This intuition holds in some but not all cases. It can be shown[43] for example, that if *R* and *U* are binary and affect *S* in the same direction but do not interact in their effects on *S*, and if *U* and *Y* are positively correlated then in Figure 6, we would have negative association between *R* and *Y*. If in a crude comparison between *R* and *Y* we found positive association (eg, if black individuals had a higher rate of an adverse outcome *Y*), then we would have an evidence of a causal relationship between *R* and *Y*, because if this were not there, the association, due to the selection bias, should be negative. In such cases, the observed associations may prove conservative estimates of the actual causal racial inequality measure under either the stronger or the weaker interpretations above. As an example, Evans et al[44] considered racial differences in the proportion of asthmatic children with severe asthma exacerbations requiring urgent medical attention in the last 12 months and found after adjusting for age, sex, and family SES that the rates of black children were 69% versus 56% for white children ($P = 0.04$). The analysis was done with a selected population, children with asthma, and the likelihood of asthma itself may of course vary across racial groups, thereby potentially distorting the associations. However, a common cause *U* (eg, moldy environment) of asthma and having an exacerbation would likely affect both in the same direction; if being black likewise increased the likelihood of asthma, then by the reasoning above we might think that this association between race and asthma exacerbations may be conservative.

However, even these approaches and arguments apply only to overall associations between race and the health outcomes. When we further adjust for adult SES, these issues of selection bias persist, possibly in more severe forms and developing approaches to handle such settings merits further research. Interestingly, however, if we are in the setting of controlling for adult SES and this adult SES measure is measured concurrently with, or following, the event S defining the subpopulation, then the weaker interpretation of the race coefficient involving direct and indirect effect racial disparity measures is still applicable. This is because, in this setting, our intervention variable, adult SES, occurs after the variable S defining the sub-population, and conditioning on S would thus not induce selection bias in the effect of adult SES. This weaker interpretation involving direct and indirect effect measures does, however, still assume that the effect of adult SES on the outcome is itself unconfounded so that associations between the adult SES measure and the outcome reflect the actual causal effects of the adult SES measure.