On the causal interpretation of race in regressions adjusting for confounding and mediating variables

Tyler J. VanderWeele[*]   Whitney Robinson[†]

[*]Harvard University, tvanderw@hsph.harvard.edu

[†]University of North Carolina Chapel Hill, whitney_robinson@unc.edu

On the causal interpretation of race in regressions adjusting for confounding and mediating variables

Tyler J. VanderWeele and Whitney Robinson

Abstract

We consider different 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 disparity 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 disparity can be decomposed into the portion that would be eliminated by equalizing adult socioeconomic status across racial groups and the portion of the disparity that would remain even if adult socioeconomic status across racial groups were equalized. We also discuss a stronger interpretation of the "effect of race" involving the joint effects of skin color, parental skin color, 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.

Introduction

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] Typically numerous other socio-demographic, economic, biological or psycho-social variables are also included in these regressions. Some of these variables may potentially be thought of as being on the pathway between race/ethnicity and whatever health outcome is under study. Other of these variables may be strongly associated with, but seemingly in no sense "caused by", race/ethnicity. The regression coefficient for race/ethnicity is interpreted as a "health disparity" irrespective of the other variables for which control has been made. However, as we will argue in this paper, how the disparity is to be interpreted depends critically on issues of temporal ordering and causality.

There have been numerous discussions of different 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 paper we offer a tentative proposal with regard to the general interpretation of a race/ethnicity variable in a regression and how this might vary given the other variables for which control has been made. What we propose certainly does not capture all of the subtleties of race/ethnicity in health disparities research but we hope it is a step in the right direction in encouraging more careful thought in what to include and what not to include in regression models that may involve race.

Part of the challenge with regard to trying to interpret race coefficients causally is that, in the formal causal inference literature, causal effects are often defined in terms of counterfactual or potential outcomes and these counterfactual or potential outcomes are in turn defined as the outcomes that would result under hypothetical interventions on the exposure.[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 disparity 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 will see how the overall disparity can be decomposed into the portion that would be eliminated by equalizing adult socioeconomic status across racial groups and the portion of the disparity 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 disparity could be eliminated by intervening on a different variable, namely socioeconomic status, which is more manipulable than race. We discuss the possibility of and the challenges with stronger interpretations of race coefficients in regression models.

The elimination of health disparities is one of the U.S. federal government's leading health objectives.[24] Persistently poorer health outcomes for some population groups may indicate violations of U.S. 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 elimination and we hope that the methodological approach in this paper will contribute to that end.

Race/Ethnicity: Correlates and Components

A racial disparity 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 "disparity" more liberally to indicate any difference, regardless of its modifiability or fairness. Such a disparity 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. The disparity itself could be assessed by comparing sample means (or some other summary measures) across two or more racial groups. To say that there is a disparity then is simply to indicate that race and the health outcome are correlated in the population under study.

If we want, however, to discuss the "effects of race," 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: it may vary depending on what is meant by race. For some, it may include skin color and its perception by others, parental skin color and its perception by others, or genetic background, say, considered separately or jointly.

Therefore, when the "effect of race" is under discussion it will generally be important to clarify what precisely is being considered. However, even then, precisely defining and assessing such "race effects" is difficult. Because race is not randomized, whether we consider skin color, parental skin color, or genetic background, singly or jointly, all of 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 so as to allow one to estimate the effects of a single genetic variant. In other contexts, if we were interested in assessing the race as an indicator of discrimination, we might be able to define the exposure of interest to be, for example, 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 be able to conduct a randomized trial which would identify the "effect of race" as more broadly conceived. If "race/ethnicity" is put in a regression, this will likely capture the effects of perceived race

along with various other factors such as neighborhood income, quality of schools, etc. that are correlated with skin color, parental skin color, genetic background.

There has 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 counterfactual or potential outcomes.[10,11] Often in the causal inference literature the position is taken that it is only meaningful to speak of a contrast of potential or counterfactual outcomes 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 non-manipulable 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 in fact correspond to a manipulable variable such as in the examples above of job application audit studies.

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

Interpretation of Race/Ethnicity in Regressions: Control for Non-Mediating Variables

To simplify discussion further we will assume that only two racial groups are under consideration (e.g. black and white), though similar remarks could apply to other comparisons. If multiple racial groups were of interest, the methods in this paper could be applied by comparing various racial groups to a single common reference racial group (e.g. comparing Asian to white and also comparing black to white).

In trying to understand health disparities, we might in principle distinguish between forward or "directed pathways" from skin color, parental skin color, or genetic background to the outcome of interest and what we might call "backdoor pathways."[26] More formally, the forward or directed pathways from skin color, parental skin color, or genetic background to the outcome are 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 which begin with an arrow pointing into one of 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, of neighborhood income, and so forth. 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 to illustrate some of the issues concerning interpretation. For now we assume all variables - skin color (SC), parental skin color (PC), genetic background (G), family/parental socioeconomic status (SES_0), 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 between the individual's skin color, parental skin color, and genetic background with the family and neighborhood socioeconomic status into which they were born. We let Y denote the subsequent health outcome of interest. 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, as we will see, it is important to clarify what is under discussion when the "effect of race" is being considered before representing it on the diagram.

We use "skin color" from this point onwards in a metaphorical sense as a generic catch-all to include all physical correlates Black versus White race in the US, such as hair texture, etc. that might be perceived by the individual or by others. The "effects of skin color" will include biologic effects of skin color (e.g. darker skin protecting against ultraviolet light), the person's understanding of her skin color and how this affects her identity and health behaviors but also, importantly, it moreover will include how others react to the person's skin color, e.g. discrimination or feelings of affinity. Objections are sometimes raised to notions such as "an effect of race" or an "effect of skin color" in 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 will here be using expressions like "effects of skin color" in the more technical sense associated with causal diagrams.[26] The arrow from perceived race to an outcome indicates some causal chain from skin color 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 skin color to the outcome.

As represented in the diagram, parental skin color may affect the individual's subsequent outcome through pathways other than through the individual's own skin color as, for example, might arise if the parents' skin color led to others discriminating against the individual as a child. Skin color and parental skin color do not of course vary independently. In most populations and with most measures of skin color the two would coincide, though exceptions can arise with parents of mixed races, adoptions, and albinism for instance. For simplicity, we will assume that the study population only has parents of a common race/ethnicity and that skin color and parental skin color do in fact coincide. If the groups constituting different "mixed race" categories (e.g. Black and Asian parents, say) were sufficiently large then these could themselves be defined as distinct racial groups. The skin color 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 (e.g. through discrimination). However, we will denote by the arrow from PC

to Y the effects of parental skin color from the time of conception onwards on the outcome. The effects prior to conception of parental skin color on the outcome, e.g. through family and neighborhood SES at the time of conception, will be captured by H.

If we put race/ethnicity into a regression the interpretation of the coefficient would likely be some combination of the effects of skin color, parental skin color, 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 skin color, parental skin color and genetic background of the individual. The task then would be to control for other variables that were correlated with skin color, parental skin color, genetic background, and the outcome but not themselves affected by race i.e. 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 prior to conception, but not post-conception. Things occurring post-conception could all 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 (i) affected the outcome of interest and (ii) were correlated with skin color, parental skin color, 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 skin color, parental skin color and genetic background considered jointly. If there were additional variables satisfying (i) and (ii), we would want to control for them as well in order to isolate the joint effects of skin color, parental skin color, and genetic background. For example, suppose some aspect of the cultural context (C) were correlated with skin color 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 skin color to cultural context. If we wanted to capture the joint effects of skin color, parental skin color, 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 skin color.

Of course we may conceive of the "effects of race" as including those aspects of the cultural context associated with skin color 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. Indeed cultural context might even be conceived of as being on the pathway from skin color insofar as skin color may predispose an individual towards certain preexisting cultural contexts. If so, we might include an arrow from skin color to cultural context. If this were the case, without adjusting for cultural context, we would be assessing the joint effects of skin color, parental skin color, genetic background, and cultural context. If we did adjust for cultural context we would have the effects of skin color, parental skin color, genetic background not through cultural context. In practice, it is unlikely 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, in Figure 2, if we wanted to capture in the "effect of race"

the joint effects of skin color, parental skin color, genetic background, and cultural context, we could replace these by our race variable R with the resulting diagram being that given in 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 would regress the outcome on our race variable (e.g. an indicator for black versus white) and also neighborhood and family SES and 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 skin color, parental skin color, genetic background and cultural context.

If desired, we might likewise not control for neighborhood SES or even family SES in regressions with race as a covariate and thereby also allow the race variable to pick up correlations with 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. We could also potentially consider several regressions, each with different controls, and each capturing or attempting to isolate different combinations of the aspects of race. However, again, if what were desired in Figure 2 were the "effects of race" conceived of as the joint effects of skin color, parental skin color, genetic background, and cultural context then controlling for all backward paths from these four variables to the health outcome of interest would arguably be the appropriate analysis.

Formalizing the Interpretation

This still leaves open the question, however, of what is the interpretation of a race/ethnicity coefficient in a regression with a specific set of control variables. We will consider two interpretations of varying strengths. The first is a stronger interpretation but one which in many cases may be implausible, and so our focus in the paper will be on the second. Suppose that one were willing to conceive of interventions of skin color, parental skin color, genetic background, and cultural context and that we were 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 skin color, parental skin color, genetic background, or cultural context) constituted a causal diagram in that there were no further backdoor pathways from skin color, parental skin color, genetic background, or cultural context through H to the outcome Y except through variables for which control had been made (e.g. family and neighborhood SES). More specifically, suppose that (i) the race variable is unassociated with Y after controlling for skin color, parental skin color, genetic background, cultural context and family and neighborhood SES and (ii) potential associations of skin color, parental skin color, 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 for greater formality). It is argued in the Appendix 1 that, under these assumptions, the race coefficient in the regression could be interpreted as the expected difference in health outcomes, for someone with a particular family and neighborhood SES, between setting skin color, parental skin color, genetic background, and cultural context to their values from a random draw from the

distribution in the white population versus settings these same variables to their values from a random draw from the distribution in the black population. See VanderWeele and Hernán[19] for further discussion of a stronger interpretation of a race coefficient in a regression. The interpretation is of course problematic in that it may be difficult to conceive of hypothetical interventions on skin color, parental skin color, genetic background, and cultural context.

If an investigator objects to the notion of skin color, parental skin color, genetic background, and cultural context being hypothetically manipulable, then, importantly, a weaker interpretation of an adjusted race coefficient is still possible. It is this weaker interpretation we will focus on in this paper. It is also argued in the 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 disparity 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 (e.g. by setting SES for each black individual 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 disparity 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. We will focus in this paper on the weaker interpretation. Note, however, that both interpretations do require that the coefficients for family and neighborhood SES correspond to the effects of these variables on the outcome. In some context 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 e.g. in a particular study in which income disparities were large, if all of the lower SES persons were black and all of 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.

Interpretation of Race/Ethnicity in Regressions: Control for Mediating Variables
In health disparities research it is also 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 describe above, such as perceived race or

genetics, these factors temporally occur after race. These factors might then be mediators of the effect of race i.e. variables on the forward pathway from race to the outcome.

Controlling for mediating factors changes the interpretation of regression coefficients and purported effect estimates. The role of socioeconomic status later in life is arguably quite distinct, from an interpretative perspective, from that in childhood or at birth. Again, an individual's socioeconomic status later in life is arguably on the pathway from skin color, parental skin, 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 skin color, parental skin color, 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.

On the other hand, control for SES later in life is perhaps sometimes done so as to assess the extent to which health disparities across racial groups are in fact 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 say. Suppose we were once again interested in the joint effects of skin color, parental skin color, 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 paths) and the extent to which they were through other pathways (the red paths). If we wanted to capture the "effects of race" conceived of as the joint effects of skin color, parental skin color, 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 give us something that we could interpret as an overall effect of skin color, parental skin color, 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 essentially want to estimate the "direct effects" of skin color, parental skin color, 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 require that there be baseline control for exposure-outcome, mediator-outcome, and exposure-mediator confounders.[26,37] However, the application of this literature to the health disparities context is potentially problematic because the "effects of race" are not generally well defined.[22] Moreover, even when effects are well defined confounders of the mediator-outcome relationship can lead to substantial biases in these effects.[26,27,31]

As before, we could potentially proceed in one of two ways. Under a stronger interpretation in which the "effects of race" were conceived of as the joint effects of skin color, parental skin color, 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 being able to conceive of counterfactuals concerning

setting skin color, parental skin color, genetic background, and cultural context to specific values, which may not be plausible and will not be pursued further here. An alternative weaker and perhaps more plausible interpretation within the context of health disparities research, however, once again 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 can be employed 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 associations between adult SES and the outcome actually reflect the effects of adult SES on the outcome; this is essentially an analogue of the mediator-outcome confounding control assumption in the literature on direct and indirect effects (no analogue of the other assumptions are necessary here because we are not intervening on the exposure, cf. Appendix). It is argued in the Appendix 1 that if these assumptions hold, 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 health disparity that would remain for individuals 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 (e.g. 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 disparity measure" not through adult SES (i.e. how much of the disparity remains after accounting for adult SES). It is also argued 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 disparity measure" through adult SES (i.e. how much of the disparity is due to difference in adult SES). It is moreover shown that the overall health disparity for those with a particular early family and neighborhood SES level is equal to the sum of these "direct" and "mediated" disparity measures. We again can interpret coefficients in this way without having to define potential outcomes with respect to race or 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.

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) 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.[30] On the odds ratios scale for logistic regression the overall disparity measure will decompose into a product (rather than the sum) of the "direct" and "mediated" disparity measures. As noted above, the interpretation of the direct and indirect effect measures given above 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 analogue of the mediator-outcome confounding control assumption in the causal inference literature on direct and indirect effects.

The methods for direct and indirect effects from the causal inference literature[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 e.g. if the effects of adult SES differed by racial groups. And indeed there is some theory and empirical evidence for such interaction between race and SES for at least some health outcomes.[38,39] When using these newer approaches that can obtain direct and mediated effect estimates even in the presence of interaction, the interpretation of these direct and mediated effect estimates would again be that given above concerning the disparity that would remain if adult SES distribution across racial groups were equalized. 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, etc. if these are likewise thought to be present.

Illustration

We provide a simple illustrate, not intended to be a full rigorous analysis, of the approaches described above with an example concerning black-white racial difference 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 U.S. 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 female. We analyzed data from non-Hispanic White and Black women who completed the 2008 follow-up visit. Respondents were aged 24 to 32 years. 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]

All analyses controlled for age. Models also were fit controlling for (i) measures of childhood family SES, childhood neighborhood SES, (ii) adult SES, (iii) and the interaction of race and childhood family SES. Childhood family SES was defined by continuous maternal education, self-reported by the respondent's biological or adoptive mom when the respondent was in secondary school . Childhood neighborhood SES was defined from the U.S. census, by the percent of adults aged 25 years or older who had completed college 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).

Finally, all models were weighted to account for Add Health's complex survey sampling and non-response.[40]

The overall excess of BMI in black versus white women was 3.74 BMI units (95% CI: 2.90, 4.58). When control was made for childhood SES (measured by years of maternal education) this difference became 3.54 (95% CI: 2.41, 4.36). When adjustment was further made for early neighborhood SES (measured by percent of adults with college degrees), this became 3.20 (95% CI: 1.65, 3.99). Under a stronger interpretation, this difference of 3.20 BMI units could be interpreted as the effects of skin color and parental skin color, genetic background, and cultural context if we thought we had adequately adjusted for confounding for the effects of skin color, parental skin color, genetic background, and cultural context; though in this illustration this seems unlikely given 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 disparity we would observe had we set early family and neighborhood SES distribution (or our measures of these variables) in the black female population to be what it was among white women. When adjustment is also made for adult SES (measured by adult education) the difference is only attenuated slightly to 3.17 (95% CI: 2.38, 3.96). Here, ignoring potential interaction between race and adult SES, the "direct effect" disparity measure is 3.17 and the "mediated effect" disparity measure (through adult education) is only 0.03 (95% CI: -0.08, 0.14). From this data, it appears that only about 1% of the BMI disparity would be eliminated if adult SES distributions were equalized. Most of the disparity does not seem to be due to difference in adult SES. When allowing for interaction between race and adult education, the estimates remain virtually unchanged. Although some of the initial disparity is explained away by these measures of neighborhood and family SES in childhood, very little of it is explained away or mediated by years of education attained in adulthood.

Discussion

In this paper we have considered the causal interpretations of the race coefficient in regressions controlling for confounding and mediating variables and have provided interpretations of these coefficients which do not require defining potential outcomes on race itself. The interpretation provided is as the disparity 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 e.g. skin color, parental skin color, genetic background, cultural context, etc. This interpretation was accomplished by framing the interpretation around intervention 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, components e.g. skin color, parental skin color, genetic background, cultural context, etc. , was thought possible, but we noted that such interventions would generally be difficult to conceive.

Our discussion here has focused on differences in outcomes across racial groups. Sometimes such differences are examined for selected populations, e.g. racial disparities for pregnant women or disparities 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 Appendix 2.

A similar approach to that proposed here could also be potentially used with factors other than socioeconomic status that may differ across racial groups. The approach could potentially also be used with other non-manipulable exposure such as sex, rather than race. Importantly, we have shown that the interpretation of the race coefficient differs substantially depending on whether variables like individual and neighborhood socioeconomic status are controlled for at birth and/or later in life. An investigator moreover need not restrict attention to only one of these analyses but may run a series of regressions, or employ modern methods for direct and indirect effect accounting also for interaction between race and socioeconomic status, to gain insight into the sources of disparities.

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:e674-681.

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 1274-1276).

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 (2003). Unequal treatment: Confronting racial and ethnic disparities in health care. Board on Health Sciences Policy. B. D. Smedley, A. Y. Stith and A. R. Nelson. Washington, D.C., Institute of Medicine of the National Academies.

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. *American Economic Review* 2004;94:991-1013.

8. Butler DM, Broockman DE. Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science*. 2011.

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. *Journal of Educational Psychology.* 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;95:477-82.

13. Hernán MA. Hypothetical interventions to define causal effects: afterthought or prerequisite? *Am J Epidemiol*. 2005;162:618-620. (Invited Commentary).

14. Holland P. 1986). 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;20:35.

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: an axiom, definition, assumption, or a 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. *Causal Inference: Statistical Perspectives and Application.* (C. Berzuini, P. Dawid and L. Bernardinelli, eds.). Wiley and Sons., 2012:101-113.

20. VanderWeele TJ, Hernán MA. Causal inference under multiple versions of treatment. *Journal of Causal Inference.* in press.

21. Greiner D, Rubin DB. Causal effects of perceived immutable characteristics. *The Review of Economics and Statistics*. 2011;93:775-785.

22. Kaufman .S. Epidemiologic analysis of racial/ethnic disparities: some fundamental issues and a cautionary example. *Social Science and Medicine*, 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/disparitiesAbout.aspx. Accessed September 26, 2013.

25. Williams DR, Jackson PB. Social sources of racial disparities in health. 10.1377/hlthaff.24.2.325, *Health Affairs*. 2005;24:325-334.

26. Pearl J. Direct and indirect effects. In: Proceedings of the Seventeenth Conference on Uncertainty and Artificial Intelligence. San Francisco: 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;Article 23.

29. VanderWeele TJ, Vansteelandt S. Conceptual issues concerning mediation, interventions and composition. *Stat Interface* - Special Issue on Mental Health and Social Behavioral Science, 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*. in press.

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*. in press.

37. Shpitser I, VanderWeele TJ. A complete graphical criterion for the adjustment formula in mediation analysis. *Int J Biostat*. 2011;7 Article 16:1-24.

38. Sanchez-Vaznaugh EV, Kawachi I, Subramanian SV, Sanchez 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. from http://www.cpc.unc.edu/projects/addhealth/design. Accessed April 4, 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, for the CHIRAH Study Group. Ethnic disparities in asthma morbidity in Chicago. *Journal of 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=(SC, PS, G, C) denote the collection of skin color, parental skin color, genetic background variables, and cultural context variables. 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 (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:

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

For the weaker interpretation, let G(0) denote a random draw of early family and neighborhood SES (i.e. 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 $pr(G(0)=x) = pr(x|R=0)$ and also because $G(0)$ is random, $pr(G(0)=x) = pr(G(0)=x|R=1)$. If the effects of family and neighborhood SES on the outcome are unconfounded conditional on R, i.e. $E[Y_x |R=1] = E[Y|R=1, x]$, 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:

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

If we sum this over the distribution of $pr(x|R=0)$ we get

$\beta_1 = \sum_x E[Y|R=1,x] \, pr(x|R=0) \, - E[Y|R=0,x] \, pr(x|R=0)$

$\beta_1 = \sum_x E[Y|R=1,x] \, pr(x|R=0) \, - E[Y|R=0]$

$\beta_1 = \sum_x E[Y_x |R=1] \, pr(G(0)=x) \, - E[Y|R=0]$

$\beta_1 = \sum_x E[Y_x |R=1, G(0)=x] \, pr(G(0)=x|R=1)- E[Y|R=0]$

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

Thus the race coefficient in the regression could be interpreted as the health disparity 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 skin color, parental skin color, 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, i.e. 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:

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

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 skin color, parental skin color, 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=(SC, PS, G, C) denote the collection of skin color, parental skin color, 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 $E[Y_{H_x(0)} |R=1, x]$ 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 white population with early family and neighborhood SES of x. Note that $pr(H_x(0)=m|x,r) = pr(H_x(0)=m) = pr(m|R=0,x)$. If the effects of M on Y are unconfounded conditional on (R,X), i.e. $E[Y_m |R=1, x] = E[Y|R=1, m, x]$, so that the associations between adult SES and the outcome reflect the actual effects of adult SES, then 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:

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

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 health disparity 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 $E[Y_{H_x(1)} |R=1, x]$ 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), i.e. $E[Y_m |R=1, x] = E[Y|R=1, m, x]$, so that the associations between adult SES and the outcome reflect the actual effects of adult SES then we have:

$\sum_m E[Y|R=1, m, x] \, pr(m | R=1, x) - \sum_m E[Y|R=1, m, x] \, pr(m | R=0, x)$
$= \sum_m E[Y_m |R=1, H_x(1)=m, x] \, pr(H_x(1)=m | R =1, x)$
$\qquad - \sum_m E[Y_m |R=1, H_x(0)=m, x] \, pr(H_x(0)=m | R=1, x)$
$= E[Y_{H_x(1)} |R=1, x] - E[Y_{H_x(0)} |R=1, x]$

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 disparity measure for those with early family and neighborhood SES of x is given by:

$E[Y|R=1, x] - E[Y|R=1, x]$
$= \sum_m E[Y|R=1, m, x]\, pr\,(m \mid R=1, x) - \sum_m E[Y|R=0, m, x]\, pr\,(m \mid R=0, x)$
$= \sum_m E[Y|R=1, m, x]\, pr\,(m \mid R=1, x) - \sum_m E[Y|R=1, m, x]\, pr\,(m \mid R=0, x)$
$\qquad + \sum_m E[Y|R=1, m, x]\, pr\,(m \mid R=0, x) - \sum_m E[Y|R=0, m, x]\, pr\,(m \mid R=0, x)$
$= \{\, E[Y_{Hx(1)} |R=1, x] - E[Y_{Hx(0)} |R=1, x] \,\} + E[Y_{Hx(0)} |R=1, x] - E[Y|R=0, m, x]$

where the second equality is obtained by adding and subtracting $\sum_m E[Y|R=1, m, x]\, pr\,(m \mid R=0, x)$ and, in the third equality, the two 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]

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. In is not infrequent, however, to also consider health disparities among selected populations. For example, racial disparities might be examined for birth outcomes for pregnant women, or 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 (e.g. 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 (e.g. skin color, parental skin color, 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 skin color, or parental skin color, or 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 skin color, parental skin color, 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 (e.g. 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 (e.g. none of skin color, parental skin color, 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 (i) the "exposure" variable R and also (ii) 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.

While 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 addressing 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 different characteristics of white and black women who become pregnant rather than to the effects of race (e.g. discrimination in response to skin color, or 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 (i) race did not in fact affect the likelihood of the event defining the population i.e. no arrows from R (or its components in Figure 5) to S or (ii) if we were able to control for common causes (e.g. 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 in the entire population, rather than in a selected population; for example, we could looks 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 (e.g. 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 (e.g. if black individuals had a higher rate of an adverse outcome Y) then we would have 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 disparity under either the stronger or weaker interpretations above. As an example, Evans et al.[44] considered racial disparities in the proportion of asthmatic children with severe asthma exacerbations requiring urgent medical attention in the last twelve months and found after adjusting for age, sex and family SES 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 (e.g. 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 handling such settings merits further research.

Figure Legends

Figure 1. Diagram illustrating relations between skin color (SC), parental skin color (PC), genetic background (G), family/parental socioeconomic status (SES_0), neighborhood socioeconomic status ($NSES_0$), history (H), and the outcome of interest Y

Figure 2. Diagram illustrating cultural context (C) which may be influence by skin color (SC)

Figure 3. Diagram with skin color, parental skin color, genetic background and culutural context replaced by a race variable (R)

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

Figure 5. Effects of race through adult SES (the blue pathways) and not through SES (the red pathways) with with skin color, parental skin color, genetic background and culutural context replaced by a race variable (R)

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)

U

H → R → S

H → SES$_0$ → S

H → NSES$_0$ → S

U → Y

6