



REVIEW ARTICLE

A new era: improving use of sociodemographic constructs in the analysis of pediatric cohort study data

Aruna Chandran¹, Emily Knapp¹, Tiange Liu¹ and Lorraine T. Dean¹

Given the diversity of sex, gender identity, race, ethnicity, and socioeconomic position (SEP) in children across the United States, it is incumbent upon pediatric and epidemiologic researchers to conduct their work in ways that promote inclusivity, understanding and reduction in inequities. Current child health research often utilizes an approach of “convenience” in how data related to these constructs are collected, categorized, and included in models; the field needs to be more systematic and thoughtful in its approach to understand how sociodemographics affect child health. We offer suggestions for improving the discourse around sex, gender identity, race, ethnicity, and SEP in child health research. We explain how analytic models should be driven by a conceptual framework grounding the choices of variables that are included in analyses, without the automatic “adjusting for” all sociodemographic constructs. We propose to leverage newly available data from large multi-cohort consortia as unique opportunities to improve the current standards for analyzing and reporting core sociodemographic constructs. Improving the characterization and interpretation of child health studies with regards to core sociodemographic constructs is critical for optimizing child health and reducing inequities in the health and well-being of all children across the United States.

Pediatric Research (2021) 90:1132–1138; <https://doi.org/10.1038/s41390-021-01386-w>

IMPACT:

- Current child health research often utilizes an approach of “convenience” in how data related to sex, race/ethnicity, and SEP are collected, categorized, and included in models.
- We offer suggestions for how scholars can improve the discourse around sex, gender identity, race, ethnicity, and SEP in child health research.
- We explain how analytic models should be driven by a conceptual framework grounding the choices of variables that are included in analyses.
- We propose to leverage newly available large cohort consortia of child health studies as opportunities to improve the current standards for analyzing and reporting core sociodemographic constructs.

INTRODUCTION

A fundamental component of both descriptive and analytic epidemiologic inquiry is a basic description of individual-level characteristics of the population under study. Identifying patterns of risk factor and disease distributions by sociodemographic factors such as sex, gender, race, ethnicity, and socioeconomic position (SEP) can be critical for the targeted distribution of limited resources.^{1–4} Understanding the effects of individual sociodemographic factors on health outcomes has been invaluable in elucidating intervention opportunities for both risk factors and diseases.^{5–9} Although nearly every published epidemiologic and public health study in the past several decades has included sociodemographic information in both descriptions and analytic models, there are few studies specifically focused on child health outcomes addressing issues with the definitions, contextualization, measurement, or appropriate use of these variables.^{10–17}

Inclusion of sociodemographic variables in child health studies can be particularly complex and problematic.⁶ The measurement and contextualization of these factors in a child are heavily influenced by the individual characteristics of the child’s parents,

siblings, caregivers, and others.^{18,19} In addition, the influences of exposures on outcomes in children are affected by stages of growth and development; consideration of the child’s life stage within his/her life course is an important part of child health inquiry.^{20,21} Finally, young children spend a majority of their time in the home, followed by a time period of significant amounts of time spent in and around school, as well as in their neighborhood. These significant changes in a child’s physical environment are accompanied by differential sociodemographic contexts that can influence a child’s health and well-being.^{22–25} Achieving consensus and consistency in the characterization of a child’s sociodemographic environment has been a challenge in child health literature.

For the past nearly 20 years, the National Institutes of Health (NIH) has required that large studies (both cohort studies as well as clinical trials) need to make data available for use by outside investigators.²⁶ In addition, in this era of “Big Data,” researchers often use secondary data made available by agencies both within and outside the health sector.²⁷ While researchers often cannot control how sociodemographic constructs within these datasets

¹Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
Correspondence: Aruna Chandran (achandr3@jhu.edu)

Received: 4 September 2020 Accepted: 30 December 2020

Published online: 18 February 2021

Table 1. Summary of recommendations for inclusion of sociodemographic factors in child health studies.

Concept	Recommendation
Conceptual framework	<ul style="list-style-type: none"> • Explicitly specify a conceptual framework. Also consider a DAG or other graphical depiction of the hypothesized relationships between variables • For sociodemographic variables including sex, gender identity, race, ethnicity, and socioeconomic position, including these constructs in your analytic model only if appropriately guided by your conceptual framework and not merely because the data are available
Sex/gender identity	<ul style="list-style-type: none"> • Be clear through a conceptual framework whether the intent is to describe participant sex or gender identity, and use the appropriate terminology • When describing the study population by sex/gender identity, be clear how the data were collected in terms of response options, who the respondent was, and when/how the data were collected • Stratify by sex/gender identity when attempting to detect differences between groups, with more than binary sex options if possible • Sex and gender identity are not proxies for one another. If according to your conceptual framework, sex and/or gender identity are confounders that should be controlled in a model, ensure you are clear on whether to control for one or both as appropriate and as the data allow
Race	<ul style="list-style-type: none"> • Race is a social categorization that has changed over time and place, not an inherent biological categorization • Do not use race as a proxy for an alternate construct such as socioeconomic position, poverty or structural racism. • Use a conceptual framework to consider how and why race is related to your outcome • Define what you believe race represents in your study and justify your choice for modeling it as a confounder effect measure modifier • In your population descriptions as well as analytic models, do not automatically reduce the data to “white” vs. “non-white” but instead carefully consider how categories might be combined based on your conceptual framework
Ethnicity	<ul style="list-style-type: none"> • Ethnicity is a categorization of shared culture and way of life, and should not be equated with race • Consider possible independent as well as intersectional effects of ethnicity and race on your chosen outcome, and use that to guide decisions about how to include race and ethnicity in your analysis
Socioeconomic position	<ul style="list-style-type: none"> • SEP is distinguished from social class or socioeconomic status in that it encompasses both material- or resource-based and prestige-based measures of socioeconomic groupings • SEP is a complex construct that cannot be represented by a single indicator such as poverty, income, or education. How to consider SEP in your analysis and which indicator(s) are most appropriate to use should be guided by your conceptual framework and your study population • In child health studies, be clear on whether your selected SEP indicator(s) represent that of the child, one parent/caregiver, or the household

were defined or collected, they still bear considerable responsibility in demonstrating the appropriate contextualization of these factors in the scholarly research that they put forth.

In this paper, we aim to summarize the major issues that have been noted to date with the use of the core and most frequently used sociodemographic variables, including sex, gender, race, ethnicity, and SEP in child health research, and put forth recommendations on how researchers can advance this field. We discuss the importance of establishing a conceptual framework for how sociodemographic factors influence the exposure(s) and outcome(s) in question, and of using this framework to decide if and how those variables should be included in an analysis. We then review the current knowledge related to the definition, analysis, and interpretation of those sociodemographic constructs. Finally, we put forth recommendations (Table 1) of how researchers can play a role in filling long-standing gaps in our understanding of how to best incorporate sociodemographic constructs in child health research.

CONCEPTUAL FRAMEWORKS

There are well-documented associations between sociodemographic characteristics and health outcomes. Race, ethnicity, gender, and socioeconomic status are well-studied examples. However, less is known about *why* these factors are salient for health or *how* these factors work together to frame health.^{13,28} Using conceptual frameworks can elucidate the role of sociodemographic factors in analyses of child health outcomes. Conceptual frameworks guide what variables are included in analytic models, and how these variables are included. Conceptual frameworks illustrate the theorized relationships between the sociodemographic and other variables in an analysis. They summarize a researcher’s understanding of previous knowledge

and application of theory as it relates to a specific topic area or research questions. An effective conceptual framework conveys the scope, levels, and key constructs of interest. They are useful tools for visually communicating complex areas of research and guide analytical decision-making. In practice, conceptual frameworks vary from broad, all-encompassing visualizations of entire research fields or theoretical frameworks to specific illustrations of a single research project.

Unfortunately, conceptual frameworks are often not explicitly stated in epidemiologic research, and sociodemographic variables are often “adjusted for” without a thought about their role and relationship to other variables.^{29,30} Factors such as race, ethnicity, sex, and SEP are often included without consideration of whether these factors are confounders or decedents of confounders, mediators, or moderators. Each of these types of variables requires different analytic treatment which may spuriously affect inferences if not handled correctly.^{31,32} For example, from a conceptual standpoint, adjusting for, or holding everything equal, except for one’s self-reported race becomes meaningless in a racialized society in which race shapes access to all aspects of a person’s life and health.^{33,34} From a statistical perspective, including a binary indicator of race in a regression analysis does not allow researchers to make inferences about differences in the exposure–outcome relationship across racial groups, illustrating the need for careful consideration of the research question, analysis, and interpretation.^{35–37}

SEX AND GENDER

The terms “sex” and “gender” refer to two distinct but interrelated concepts that are recognized globally as the core social determinants of health across a wide variety of geographic settings.^{38,39} There has been fairly extensive discussions in the

literature aiming to clarify these as two distinct concepts and calling for the need to standardize the use of these terms.^{40–44} Traditionally in the literature, the terms sex and gender have been conflated to represent a combined construct of biological characteristics and cultural expression, and the categorization has been binary despite evidence that there are more than two sexes.^{45,46} However, by definition, sex refers to the set of biological attributes in humans and animals that are associated with specific physical and physiological features, including relevant chromosomes, gene expression, hormone function, and reproductive/sexual anatomy.³⁹ Gender refers to the set of cultural meanings ascribed to or associated with patterns of behavior, experience, and personality that are labeled as feminine or masculine.⁴⁷ Sex and gender have separate although often interactive and synergistic effects on health, illness, well-being, and experience with the health care system.⁴⁴ Therefore, they are inappropriate proxies for one another, and should be measured and analyzed distinctly.⁴⁸ It is also important to note that sex and gender are different from conceptualization of sexual orientation or sexual attraction. Although the measurement and analysis of sexual orientation, which describes romantic or sexual attraction, is not discussed in this paper, it is important to consider within the context of child health research.

Many experts have recommended that in accordance with its strictest definition, sex is conceptualized as a binary factor, and the terms “male” and “female” should be used in its description.⁴⁰ However, this has been called into question as more research has been done into intersex conditions, grouped under the term Disorders of Sex Development, which have a prevalence of up to 1 per 100 persons.⁴⁹ There are more than 20 conditions included in the categorization of Disorders of Sex Development, the more commonly known of which include Congenital Adrenal Hyperplasia, Gonadal Dysgenesis, Androgen Insensitivity Syndrome, Turner Syndrome, and Klinefelter Syndrome. These individuals would not fit into a binary definition of sex, and therefore sex likely requires a third category in research work; studies that only capture sex as “male” or “female” would fail to accurately capture or represent intersex persons.⁵⁰

From a methodologic standpoint, there have been some published recommendations regarding the inclusion of sex in the analysis of research findings.^{46,51} In 2016, National Institutes of Health (NIH) released the Sex as a Biological Variable (SABV) policy, which states that “NIH expects that SABV will be factored into research designs, analyses, and reporting in vertebrate animal and human studies.”⁴⁵ The SABV policy suggests that regardless of whether the study was powered to detect sex differences, data should be disaggregated to explore any differences that could be obscured when data from males and females are pooled, and therefore that key relationships between the exposure and outcome should be analyzed for males and females separately.⁴⁵ Researchers have noted that when sex is included in models in most of the epidemiologic literature, it is for the most part treated as a confounder, thus neglecting its potential role as an effect measure modifier.^{52–54} Importantly, investigators should use their conceptual framework to determine what about sex differences is important in the analysis. If there are underlying characteristics that traditionally differ by sex, then those should be measured and analyzed directly without using sex as a loose proxy.⁵⁵

In contrast, gender is more commonly recognized as a multidimensional construct that includes gender identity, gender expression, and gender label (applying a name and definition to one’s gender identity and expression).¹⁰ Gender identity, according to the Institute of Medicine’s 2011 report, “refers to a person’s basic sense of being a man or boy, a woman or girl, or another gender (e.g., transgender, bigender, or gender queer—a rejection of the traditional binary classification of gender).”⁴⁷ Gender expression “denotes the manifestation of characteristics in one’s personality, appearance, and behavior that are culturally defined

as masculine or feminine.”⁴⁷ The most commonly accepted and used construct in both measurement and analysis in research is gender identity, generally referring to how an individual perceives their own gender.¹⁰

The naming or categorization of gender identity remains inconsistent in the literature. Gender minority is an “umbrella” term that refers to transgender- and gender-nonconforming people, that is, people whose current gender identity or gender expression do not conform to social expectations based on their sex assigned at birth.⁴⁷ Studies suggest that gender-typical as well as transgender children as young as age 3 years can reliably identify their gender.⁵⁶ The Center for Disease Control and Prevention’s Youth Risk Behavior Surveillance System, the Center of Excellence for Transgender Health, and the US Department of Education’s School Climate Survey each conduct measures of gender identity among adolescents and youth. Each of these categorize gender identity similarly, using “man,” “woman,” “transgender man,” “transgender woman,” “gender nonconforming,” and “other.”^{10,50} However, given the rapid evolution of awareness, knowledge and exposure in society, recommendations on appropriate and acceptable terminology continue to evolve.⁵⁷

There is relatively little published guidance related to the appropriate inclusion of gender identity in analytic models. The fundamental question to be asked is based on a specific research question, and which construct (sex, gender identity, or both) most appropriately measures and characterizes what the question aims to answer. Nowatzki and Grant⁴³ argue that sex disaggregation alone is insufficient to understand gender-based contexts of health services, because it implies that differences in social, political, and economic power between individuals of different gender identities, and the health consequences of those inequalities, are not addressed. They concluded that regardless of the methodological approach taken, it is possible to do both a sex- and gender-based analysis, provided that appropriate indicators are incorporated into the data collection instruments.⁴³ Questions remain regarding how both of these variables can be used in the same model, given the collinearity between the two constructs.^{46,51,55}

In summary, although there is growing recognition of the need to separate constructs of sex and gender in epidemiologic inquiry and some recommendations for the importance of including sex differences in analytic models, there remain several open questions and inconsistencies in how to define and categorize sex as well as gender identity and how to appropriately incorporate both of these constructs in child health research.

RACE AND ETHNICITY

Race and ethnicity are now widely acknowledged as two rapidly evolving and poorly defined constructs; however, this was not always the case. The term “race” was first used to refer to genetically distinct groups within a species. However, our current-day uses of the term race do not reflect genetically distinct groups, but instead focus on taxonomies of human groups based on perceived physical characteristics and geographic origin.⁵⁸ Race, as currently conceived, is a poorly defined marker for biologic and genetic variation found across all humans, as there is greater genetic variation within racial groups than across racial groups.^{9,59–61} If interested in groups that are genetically similar, examining genetic ancestry is more appropriate, as it is based on populations that are geographically, culturally, and linguistically similar over time; however, groups by genetic ancestry are not equivalent to the socially and politically designated race groups.^{62,63}

Ethnicity is used to classify human populations based on shared culture and way of life, especially as reflected in language, folklore, religious and other institutional forums, material objects such as clothing and food, and cultural products such as music, literature, and art.⁹ Although race and ethnicity have different meanings, the

conceptual confusion between them within the research literature emerged in as early as 1978 when the US Office of Management and Budget created "race/ethnicity" as a combined category in the reporting of federal statistics.⁶⁴ In the epidemiologic literature, the two terms are often used interchangeably, a combined expression of "race/ethnicity" is often included in analytic models, and the terms are rarely precisely defined or described by researchers.^{65–68}

The epidemiologic literature also reflects the fluid and ill-defined categorization of race and ethnicity. Related to child health, natality statistics from the National Center for Health Statistics prior to 1989 reported the race of a newborn based on the race of both parents. However, when parents were of different races and one parent was white, the child was assigned the race of the non-white parent.⁶⁹ These practices were rooted on the "one-drop rule" (Laws in the 1700s through the twentieth century, and held up by courts as late as 1985, which criminalized interracial marriages and designated White person as one "who has no trace whatsoever of any blood other than Caucasian" and took a "fractional, blood-borne approach" to define who was Black.) that reinforced white superiority and that being assigned to a white race was a privilege only for those of solely white generational lines. Since 1989, the race of the newborn is based on the race of the mother alone.^{70,71} In another example, Comstock et al.⁶⁷ reported in their review of articles published in *the American Journal of Epidemiology* and *the American Journal of Public Health* from 1996 to 1999 that the number of categories for race and/or ethnicity in the literature ranged from 0 to 24, with an average of 3.14.⁶⁷ In another extreme, Flores¹¹ showed in a review of studies exploring racial/ethnic disparities in the health and health care of US children that combining all non-white children into one group occurred in 9% of the 122 studies excluded from their final analysis.¹¹ Researchers often choose to combine or split certain categories, either based on the granularity of information available or to ensure adequate statistical power.^{11,67} Importantly, the majority of studies fail to explain or justify how race and/or ethnicity information was collected or combined, thus obscuring the process to readers.^{1,66–68,72}

Studies have shown that race and/or ethnicity are often conceptualized as proxy measures for other concepts that are known or believed to be correlated with them (i.e., poverty, discrimination, cultural factors, structural racism, or unspecified biological differences).^{14,73,74} Walsh and Ross¹⁴ showed that in articles published in three general pediatric journals (*Pediatrics*, *Journal of Pediatrics*, and *Archives of Pediatrics and Adolescent Medicine*) between July 1999 through June 2000, 35% of the articles that reported race and/or ethnicity data did not report any socioeconomic information (40/115) and 24% that discussed race and/or ethnicity did not discuss socioeconomic factors (11/45), leading the authors to conclude that researchers are using race and/or ethnicity as an explanatory variable to represent poverty.¹⁴ Race is often included in clinical algorithms with no description of why racial differences in outcomes may exist, despite the inherently casual interpretation of such algorithms. If racism, socioeconomic differences, or other societal factors explain the differences in clinical outcomes, including race in such predictive models may actually increase disparities in health outcomes.⁷⁵

In 1993, the Centers for Disease Control and Prevention recommended researchers to clearly indicate their reason(s) for analyzing data on race and ethnicity.⁷⁶ Subsequently, in 2000 and 2001, the American Academy of Pediatrics' Committee on Pediatric Research as well as the editors of the *Archives of Pediatrics and Adolescent Medicine* recommended that researchers not use race and/or ethnicity as explanatory variables in place of target underlying concepts (i.e., poverty, racism, etc.) that can and should be measured directly.^{12,77} More recently, the American Academy of Pediatrics has shifted to prioritizing the role of racism, the "system of structuring opportunity and assigning value based on the social interpretation of how one looks (which is what we

call 'race')," rather than race itself, in investigations of trends in child health.⁷⁸ Despite these recommendations, inserting race and/or ethnicity covariates continues and has, in fact, been found to be increasing in child health research.^{16,66,67}

Relatively little has been published on appropriate analytic methods for including race and/or ethnicity in models when justified by an underlying conceptual framework.⁷⁹ LaVeist suggested instead of merely "controlling" for race either to report models stratified by race groups or specify a multiplicative interaction term between the race variable and each of the other independent variables to explore more fully the effects of race in the analysis.³⁵ Interpretations of these models move us toward understanding how our exposure or interventions might operate differently in one group than another, rather than erroneously attributing differences in treatment effects to race itself. Other notable guidance offered includes Jones' recommendations for use of race in epidemiology, Kaufman and Cooper statements on valid approaches to using race in biomedical research, and VanderWeele's approaches to causal interpretations of race.^{34,79,80} As decisions about how to capture race and ethnicity continue to evolve and allow for more complex self-identification, researchers will need to be more thoughtful about how best to categorize people for analysis.

In summary, despite representing two different social constructs, race and ethnicity are often combined in epidemiologic inquiry, and frequently included in analytic models either as poor proxies for other constructs or without any justification at all. Even when appropriately justified in the conceptual framework, further research is needed as to how best to include race or ethnicity in child health research.

SOCIOECONOMIC POSITION

There are numerous terms to describe socioeconomic conditions, such as poverty, socioeconomic status, SEP, social class, social stratification, and social inequality. In general, these terms are used by researchers interchangeably, in spite of their different origins, theoretical bases and interpretations.⁸¹ For the purposes of this discussion, we will use the term SEP to refer to all of these sociologic concepts. SEP is distinguished from social class or socioeconomic status in that it encompasses both material- or resource-based and prestige-based measures of socioeconomic groupings.^{82,83} In epidemiologic studies on child health, commonly used SEP indicators include parental (mother and/or father) education and occupation, household income, wealth, poverty level, living conditions, neighborhood socioeconomic characteristics, and a variety of composite scales which consolidate multiple domains into a single construct.^{13,16,84–86} SEP is relatively frequently reported in the child health literature, and has increasingly been highlighted as an underlying determinant of a variety of child health outcomes.^{16,87}

There has been much controversy on the dimensions that can best assess SEP; SEP is widely acknowledged to be a multi-dimensional construct comprising diverse socioeconomic factors, and that different indicators are often used to describe correlated but different aspects of SEP.⁸ For example, income and wealth most directly measure material circumstances, whereas education can reflect a range of noneconomic social characteristics, including general and health-related knowledge.⁸⁸ However, over the past three decades, use of a single indicator to "control for SEP" has been commonly noted in the literature.^{89,90} For example, education is often used as a proxy for income, and income is often used as a proxy for wealth.^{13,90} Although SEP indicators have been widely assumed to be correlated, studies have indicated that these correlations are generally not strong enough to justify using one as proxy for all others.^{17,90–92} Braveman et al.⁹⁰ analyzed five nationally or state-wide-representative data sources, and reported that the income–education correlation is mostly <0.5.⁹⁰

Researchers have been recommending the use of more than one indicator to measure and represent SEP over the past several decades.^{13,91,93} Potential advantages of doing so specifically in child health research include both improving the accuracy of the measurement of the construct and allowing for a fuller understanding of the mechanistic pathways in the relationships between SEP and child health.⁹⁴

Beyond the choice of indicators, the practical use of SEP in statistical analysis has additional challenges. First, although an individual's SEP may change over time, most epidemiology research in child health relies on SEP ascertained at a single point in time.⁸ Second, children are dependent on their parents/caregivers. However, it is often unclear whose SEP characteristics and under what circumstances should be measured and assigned. For example, there is evidence that the influence of maternal and paternal education and income is actually different for certain outcomes.^{95,96} Third, how to quantify certain indicators is not clear, and certainly, geographic locale, calendar year, and individual demographics affect what level of difference SEP indicators most influence health outcomes.^{8,28,97}

In summary, there is no question that SEP affects child health and well-being. Improving our understanding of how best to characterize and analyze this construct to optimize potential interventions to improve child health is critical.

DISCUSSION

Our social, economic, and physical environments are well-recognized to influence child health, development, and well-being. Given the remarkable diversity of sex, gender identity, race, ethnicity, and SEP in children across the United States, it is incumbent upon pediatric and epidemiologic researchers to conduct their work in ways that promote inclusivity, understanding and ultimately reduction in inequities. In this paper, we underscore problems with the conceptualization, categorization, and analysis in current research in considering these core sociodemographic constructs. Current research often utilizes an approach of "convenience" in how data related to these constructs are collected, categorized, and included in models, and it is time for the field to be more systematic and thoughtful in its approach to understand how sociodemographics affect child health.

Publicly available data from large studies or consortia can be leveraged for their large sample sizes, and demographically and geographically diverse populations. Researchers have discussed the numerous benefits of promoting access to research data.^{98,99} Specific to child health, examples in the literature illustrate how accessing publicly available data can advance knowledge beyond what most smaller single cohorts could answer related to important outcomes such as obesity, mental health, and mortality.^{100–102} Entire datasets from large often nationally representative studies or surveys such as the National Survey of Children's Health and the FLASHE study are available for public use.^{103,104} Data from a consortium of child cohorts called the Environmental Influences on Child Health Outcomes will have data available in the near future.¹⁰⁵ What is missing from the literature is guidance on how the research community has an obligation to improve the discourse related to sociodemographic characteristics and disparities in ways that works to reduce inequities across all subpopulations.

Our paper has several limitations. First, we do not consider how to improve data collection or measurement of these constructs in child health research. While this article focuses on recommendations for users of data from repositories or publicly available sources, we do believe there is a need for future work discussing optimal approaches for defining, measuring, and collecting sociodemographic data in child health research. Second, there are several social characteristics that are not discussed in this paper, such as sexual orientation, immigration status, and so on.

Third, in this paper, we do not consider ways to improve multilevel research, such as how best to characterize SEP when considering the influence of one's neighborhood in their health. Although outside the scope of this discussion, we believe these are critical concepts that should be considered in the future.

We offer suggestions for how scholars can improve the discourse around sex, gender identity, race, ethnicity, and SEP in child health research. Improving the characterization and interpretation of child health studies with regards to core sociodemographic constructs is a critical component of optimizing child health and reducing inequities in the health and well-being of all children across the United States.

ACKNOWLEDGEMENTS

We thank Dr. Lisa Jacobson for her guidance and edits on this manuscript. L.T.D. is supported by grants from the National Institute of Mental Health (R25MH083620) and the National Cancer Institute (K01CA184288). This publication was supported by the Environmental Influences on Child Health Outcomes (ECHO) program, Office of the Director, National Institutes of Health, under Award Number U24OD023382 (Data Analysis Center). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

AUTHOR CONTRIBUTIONS

A.C., E.K., and T.L. contributed to the conception, design, acquisition, and interpretation of the information. L.T.D. provided critical input in the intellectual content. All authors contributed to the writing of the manuscript and approved the final version for publication.

ADDITIONAL INFORMATION

Competing interests: The authors declare no competing interests.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

1. Bhopal, R. Is research into ethnicity and health racist, unsound, or important science? *BMJ* **314**, 1751–1756 (1997).
2. James, S. A. Epidemiologic research on health disparities: some thoughts on history and current developments. *Epidemiol. Rev.* **31**, 1–6 (2009).
3. Seith, D. & Isakson, E. *Who are America's Poor Children? Examining Health Disparities among Children in the United States* (Mailman School of Public Health, 2011).
4. National Research Council and Institute of Medicine. *Children's Health, the Nation's Wealth: Assessing and Improving Child Health* (National Academies Press, 2004).
5. Sanders-Phillips, K., Settles-Reaves, B., Walker, D. & Brownlow, J. Social inequality and racial discrimination: risk factors for health disparities in children of color. *Pediatrics* **124**(Suppl. 3), S176–S186 (2009).
6. Spencer, N. Social, economic, and political determinants of child health. *Pediatrics* **112**(Part 2), 704–706 (2003).
7. Marmot, M., Friel, S., Bell, R., Houweling, T. A. & Taylor, S. Commission on Social Determinants of H. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* **372**, 1661–1669 (2008).
8. Daly, M. C., Duncan, G. J., McDonough, P. & Williams, D. R. Optimal indicators of socioeconomic status for health research. *Am. J. Public Health* **92**, 1151–1157 (2002).
9. Institute of Medicine. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare* (National Academies Press, 2003).
10. Temkin, D., Belford, J., McDaniel, T., Stratford, B. & Parris, D. Improving measurement of sexual orientation and gender identity among middle and high school students. *Child Trends* **23**, 1–2 (2017).
11. Flores, G. Committee on Pediatric R. Technical report—racial and ethnic disparities in the health and health care of children. *Pediatrics* **125**, e979–e1020 (2010).
12. American Academy of Pediatrics. Race/ethnicity, gender, socioeconomic status—research exploring their effects on child health: a subject review. *Pediatrics* **105**, 1349–1351 (2000).
13. Cheng, T. L. & Goodman, E. Committee on Pediatric R. Race, ethnicity, and socioeconomic status in research on child health. *Pediatrics* **135**, e225–e237 (2015).

14. Walsh, C. & Ross, L. F. Are minority children under- or overrepresented in pediatric research? *Pediatrics* **112**, 890–895 (2003).
15. Ross, L. F. & Walsh, C. Minority children in pediatric research. *Am. J. Law Med.* **29**, 319–336 (2003).
16. Brahan, D. & Bauchner, H. Changes in reporting of race/ethnicity, socioeconomic status, gender, and age over 10 years. *Pediatrics* **115**, e163–e166 (2005).
17. Braveman, P., Cubbin, C., Marchi, K., Egertter, S. & Chavez, G. Measuring socioeconomic status/position in studies of racial/ethnic disparities: maternal and infant health. *Public Health Rep.* **116**, 449–463 (2001).
18. Case, A. & Paxson, C. Parental behavior and child health. *Health Aff.* **21**, 164–178 (2002).
19. Kenney, M. K. Child, family, and neighborhood associations with parent and peer interactive play during early childhood. *Matern. Child Health J.* **16**(Suppl. 1), S88–S101 (2012).
20. Braveman, P. & Barclay, C. Health disparities beginning in childhood: a life-course perspective. *Pediatrics* **124**(Suppl. 3), S163–S175 (2009).
21. Kuh, D., Ben-Shlomo, Y., Lynch, J., Hallqvist, J. & Power, C. Life course epidemiology. *J. Epidemiol. Community Health* **57**, 778–783 (2003).
22. Huang, K. Y., Cheng, S. & Theise, R. School contexts as social determinants of child health: current practices and implications for future public health practice. *Public Health Rep.* **128**(Suppl. 3), 21–28 (2013).
23. Maitland, C., Stratton, G., Foster, S., Braham, R. & Rosenberg, M. A place for play? The influence of the home physical environment on children's physical activity and sedentary behaviour. *Int. J. Behav. Nutr. Phys. Act.* **10**, 99 (2013).
24. Sampson, R. J. The neighborhood context of well-being. *Perspect. Biol. Med.* **46** (Suppl.), S53–S64 (2003).
25. Christian, H. et al. The influence of the neighborhood physical environment on early child health and development: a review and call for research. *Health Place* **33**, 25–36 (2015).
26. National Institutes of Health (NIH). *Final NIH Statement on Sharing Research Data* (NIH, 2003).
27. Gandomi, A. & Haider, M. Beyond the hype: big data concepts, methods and analytics. *Int. J. Inf. Manag.* **35**, 137–144 (2015).
28. Hamilton, D. *Post-Racial Rhetoric, Racial Health Disparities, and Health Disparity Consequences of Stigma, Stress and Racism* (Washington Center for Equitable Growth, 2017).
29. Krieger, N. Epidemiology and the web of causation: has anyone seen the spider? *Soc. Sci. Med.* **39**, 887–903 (1994).
30. Krieger, N. Theories for social epidemiology in the 21st century: an ecosocial perspective. *Int. J. Epidemiol.* **30**, 668–677 (2001).
31. Krieger, N., Davey & Smith, G. The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. *Int. J. Epidemiol.* **45**, 1787–1808 (2016).
32. Hernandez-Diaz, S., Schisterman, E. F. & Hernan, M. A. The birth weight “paradox” uncovered? *Am. J. Epidemiol.* **164**, 1115–1120 (2006).
33. Galea, S. & Link, B. G. Six paths for the future of social epidemiology. *Am. J. Epidemiol.* **178**, 843–849 (2013).
34. Kaufman, J. S. & Cooper, R. S. Commentary: considerations for use of racial/ethnic classification in etiologic research. *Am. J. Epidemiol.* **154**, 291–298 (2001).
35. LaVeist, T. A. Beyond dummy variables and sample selection: what health services researchers ought to know about race as a variable. *Health Serv. Res.* **29**, 1–16 (1994).
36. Westreich, D. & Greenland, S. The table 2 fallacy: presenting and interpreting confounder and modifier coefficients. *Am. J. Epidemiol.* **177**, 292–298 (2013).
37. Kaufman, J. S. Statistics, adjusted statistics, and maladjusted statistics. *Am. J. Law Med.* **43**, 193–208 (2017).
38. Piccini, P., Montagnani, C. & deMartino, M. Gender disparity in pediatrics: a review of the current literature. *Italian J. Pediatr.* **44**, <https://doi.org/10.1186/s13052-017-0437-x> (2018).
39. Coen, S. & Banister, E. *What a Difference Sex and Gender Make: A Gender, Sex and Health Research Casebook* (Canadian Institutes of Health Research, 2012).
40. Clayton, J. A. & Tannenbaum, C. Reporting sex, gender, or both in clinical research? *JAMA* **316**, 1863–1864 (2016).
41. The Society for Women's Health Research. Institute of medicine report validates the science of sex differences. *J. Womens Health Gen. Based Med.* **10**, 303–304 (2001).
42. Day, S., Mason, R., Lagosky, S. & Rochon, P. A. Integrating and evaluating sex and gender in health research. *Health Res. Policy Syst.* **14**, 75 (2016).
43. Nowatzki, N. & Grant, K. R. Sex is not enough: the need for gender-based analysis in health research. *Health Care Women Int.* **32**, 263–277 (2011).
44. Doyal, L. Sex and gender: the challenges for epidemiologists. *Int. J. Health Serv.* **33**, 569–579 (2003).
45. Clayton, J. A. Applying the new SABV (sex as a biological variable) policy to research and clinical care. *Physiol. Behav.* **187**, 2–5 (2018).
46. Krieger, N. Genders, sexes, and health: what are the connections—and why does it matter? *Int. J. Epidemiol.* **32**, 652–657 (2003).
47. Institute of Medicine. *The Health of Lesbian, Gay, Bisexual and Transgender People: Building a Foundation for a Better Understanding* (The National Academies Press, 2011).
48. Ruiz-Cantero, M. T. et al. A framework to analyse gender bias in epidemiological research. *J. Epidemiol. Community Health* **61**(Suppl. 2), ii46–ii53 (2007).
49. Arboleda, V. A., Sandberg, D. E. & Vilain, E. DSDs: genetics, underlying pathologies and psychosexual differentiation. *Nat. Rev. Endocrinol.* **10**, 603–615 (2014).
50. Kann, L. et al. Youth risk behavior surveillance—United States, 2015. *MMWR Surveill. Summ.* **65**(Suppl. 56), 1–174 (2016).
51. Bird, C. E. & Rieker, P. P. Gender matters: an integrated model for understanding men's and women's health. *Soc. Sci. Med.* **48**, 745–755 (1999).
52. Moerman, C. & van Mens-Verhulst, J. Gender-sensitive epidemiologic research: Suggestions for a gender-sensitive approach towards problem definition, data collection and analysis in epidemiological research. *Psychol. Health Med.* **9**, 41–52 (2004).
53. Jahn, I. & Foraita, R. Gender-sensitive epidemiological data analysis: methodological aspects and empirical outcomes. Illustrated by a health reporting example. *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz* **51**, 13–27 (2008).
54. Nieuwenhoven, L. & Klinge, I. Scientific excellence in applying sex- and gender-sensitive methods in biomedical and health research. *J. Women's Health* **19**, 313–321 (2010).
55. Springer, K. W., Mager Stellman, J. & Jordan-Young, R. M. Beyond a catalogue of differences: a theoretical frame and good practice guidelines for researching sex/gender in human health. *Soc. Sci. Med.* **74**, 1817–1824 (2012).
56. Olson, K. R. & Gulgoz, S. Early findings from the TransYouth Project: gender development in transgender children. *Child Dev. Perspect.* **12**, 93–97 (2018).
57. Ainsworth, C. Sex redefined. *Nature* **518**, 288–291 (2015).
58. Krieger, N. in *Epidemiology and the People's Health*, 86–92 (Oxford Univ. Press, 2011).
59. Fullilove, M. T. Comment: abandoning “race” as a variable in public health research—an idea whose time has come. *Am. J. Public Health* **88**, 1297–1298 (1998).
60. Bhopal, R. & Donaldson, L. White, European, Western, Caucasian, or what? Inappropriate labeling in research on race, ethnicity, and health. *Am. J. Public Health* **88**, 1303–1307 (1998).
61. Yudell, M., Roberts, D., DeSalle, R. & Tishkoff, S. Science and society. Taking race out of human genetics. *Science* **351**, 564–565 (2016).
62. Fujimura, J. H. & Rajagopalan, R. Different differences: the use of ‘genetic ancestry’ versus race in biomedical human genetic research. *Soc. Stud. Sci.* **41**, 5–30 (2011).
63. Morning, A. Does genomics challenge the social construction of race? *Sociol. Theory* **32**, 189–207 (2014).
64. Office of Management and Budget. *Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity* (OMB, 2016).
65. Senior, P. A. & Bhopal, R. Ethnicity as a variable in epidemiological research. *BMJ* **309**, 327–330 (1994).
66. Moubarac, J. C. Persisting problems related to race and ethnicity in public health and epidemiology research. *Rev. Saude Publica* **47**, 104–115 (2013).
67. Comstock, R. D., Castillo, E. M. & Lindsay, S. P. Four-year review of the use of race and ethnicity in epidemiologic and public health research. *Am. J. Epidemiol.* **159**, 611–619 (2004).
68. Ahdieh, L. & Hahn, R. A. Use of the terms ‘race’, ‘ethnicity’, and ‘national origins’: a review of articles in the American Journal of Public Health, 1980-1989. *Ethn. Health* **1**, 95–98 (1996).
69. National Center for Health Statistics. *Vital Statistics of the United States: Natality* (Centers for Disease Control and Prevention, 1999).
70. Ho, A. K., Sidanius, J., Levin, D. T. & Banaji, M. R. Evidence for hypodescent and racial hierarchy in the categorization and perception of biracial individuals. *J. Pers. Soc. Psychol.* **100**, 492–506 (2011).
71. Hickman, C. B. The devil and the one drop rule: racial categories, African Americans, and the U.S. Census. *Mich. Law Rev.* **95**, 1161–1265 (1997).
72. Jones, C. P., LaVeist, T. A. & Lillie-Blanton, M. “Race” in the epidemiologic literature: an examination of the American Journal of Epidemiology, 1921–1990. *Am. J. Epidemiol.* **134**, 1079–1084 (1991).
73. Williams, D. R. Race/ethnicity and socioeconomic status: measurement and methodological issues. *Int. J. Health Serv.* **26**, 483–505 (1996).
74. Passel, J. S. Demographic and social trends affecting the health of children in the United States. *Ambul. Pediatr.* **2**(Suppl.), 169–179 (2002).
75. Vyas, D. A., Eisenstein, L. G. & Jones, D. S. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.* **383**, 874–882 (2020).

76. Centers for Disease Control and Prevention. Use of race and ethnicity in public health surveillance: Summary of the CDC/ATSDR workshop. *Morb. Mortal. Wkly. Rep.* **17**, 42 (1993).
77. Rivara, F. & Finberg, L. Use of the terms race and ethnicity. *Arch. Pediatr. Adolesc. Med.* **155**, 119 (2001).
78. Trent, M., Dooley, D. G. & Douge, J. Section On Adolescent H, Council On Community P, Committee On A. The impact of racism on child and adolescent health. *Pediatrics* **144**, e20191765 (2019).
79. VanderWeele, T. J. & Robinson, W. R. On the causal interpretation of race in regressions adjusting for confounding and mediating variables. *Epidemiology* **25**, 473–484 (2014).
80. Jones, C. P. Invited Commentary: “race,” racism, and the practice of epidemiology. *Am. J. Epidemiol.* **154**, 299–304 (2001).
81. Liberatos, P., Link, B. G. & Kelsey, J. L. The measurement of social class in epidemiology. *Epidemiol. Rev.* **10**, 87–121 (1988).
82. Krieger, N., Williams, D. R. & Moss, N. E. Measuring social class in US public health research: concepts, methodologies, and guidelines. *Annu. Rev. Public Health* **18**, 341–378 (1997).
83. Krieger, N. A glossary for social epidemiology. *J. Epidemiol. Community Health* **55**, 693–700 (2001).
84. Chittleborough, C. R., Baum, F. E., Taylor, A. W. & Hiller, J. E. A life-course approach to measuring socioeconomic position in population health surveillance systems. *J. Epidemiol. Community Health* **60**, 981–992 (2006).
85. Currie, J. Healthy, wealthy and wise: socioeconomic status, poor health in childhood, and human capital development. *J. Econ. Lit.* **47**, 87–112 (2009).
86. de Neubourg, E., Borghans, L., Coppens, K. & Jansen, M. Explaining children’s life outcomes: parental socioeconomic status, intelligence and neurocognitive factors in a dynamic life cycle model. *Child Indic. Res.* **11**, 1495–1513 (2018).
87. Laflamme, L., Hasselberg, M. & Burrows, S. 20 years of research on socioeconomic inequality and children’s unintentional injuries understanding the cause-specific evidence at hand. *Int. J. Pediatr.* **2010**, 1–23 (2010).
88. Lynch, J. W. & Kaplan, G. A. *Socioeconomic Position* (Oxford Univ. Press, 2000).
89. Galobardes, B., Lynch, J. & Smith, G. D. Measuring socioeconomic position in health research. *Br. Med. Bull.* **81–82**, 21–37 (2007).
90. Braveman, P. A. et al. Socioeconomic status in health research: one size does not fit all. *JAMA* **294**, 2879–2888 (2005).
91. Abramson, J. H., Gofin, R., Habib, J., Pridan, H. & Gofin, J. Indicators of social class. A comparative appraisal of measures for use in epidemiological studies. *Soc. Sci. Med.* **16**, 1739–1746 (1982).
92. Winkleby, M. A., Jatulis, D. E., Frank, E. & Fortmann, S. P. Socioeconomic status and health: how education, income, and occupation contribute to risk factors for cardiovascular disease. *Am. J. Public Health* **82**, 816–820 (1992).
93. Galobardes, B., Shaw, M., Lawlor, D. A., Lynch, J. W., Davey & Smith, G. Indicators of socioeconomic position (Part 1). *J. Epidemiol. Community Health* **60**, 7–12 (2006).
94. Braveman, P. & Gottlieb, L. The social determinants of health: it’s time to consider the causes of the causes. *Public Health Rep.* **129**(Suppl. 2), 19–31 (2014).
95. Vollmer, S., Bommer, C., Krishna, A., Harttgen, K. & Subramanian, S. V. The association of parental education with childhood undernutrition in low- and middle-income countries: comparing the role of paternal and maternal education. *Int. J. Epidemiol.* **46**, 312–323 (2017).
96. Lindeboom, M., Llana-Nozal, A., Van & der Klaauw, B. Parental education and child health: evidence from a schooling reform. *J. Health Econ.* **28**, 109–131 (2009).
97. Herring, C. & Henderson, L. Wealth inequality in black and white: cultural and structural sources of the racial wealth gap. *Race Soc. Probl.* **8**, 4–17 (2016).
98. Arzberger, P. et al. Promoting access to public research data for scientific, economic, and social development. *Data Sci. J.* **3**, 135–152 (2004).
99. Stockemer, D., Koehler, S. & Lentz, T. Data access, transparency, and replication: new insights from the political behavior literature. *Polit. Sci. Polit.* **51**, 799–803 (2018).
100. Hammond, R. et al. Predicting childhood obesity using electronic health records and publicly available data. *PLoS ONE* **14**, e0215571 (2019).
101. Thakrar, A. P., Forrest, A. D., Maltenfort, M. G. & Forrest, C. B. Child mortality in the US and 19 OECD comparator nations: a 50-year time-trend analysis. *Health Aff.* **37**, 140–149 (2018).
102. Blair, L. M. Publicly available data and pediatric mental health: leveraging big data to answer big questions for children. *J. Pediatr. Health Care* **30**, 84–87 (2016).
103. Nebeling, L., Dwyer, L. & Oh, A. The FLASHE Study: a publicly available data resource. *Curr. Dev. Nutr.* **4**(Suppl. 2), 1337 (2020).
104. Ghandour, R. M. et al. The Design and Implementation of the 2016 National Survey of Children’s Health. *Matern. Child Health J.* **22**, 1093–1102 (2018).
105. Gillman, M. W. & Blaisdell, C. J. Environmental influences on Child Health Outcomes, a Research Program of the National Institutes of Health. *Curr. Opin. Pediatr.* **30**, 260–262 (2018).