Analysis Considerations for Race, Ethnicity, and Gender Variables

Developed by: Northwestern University's Department of Preventive Medicine's Working Group on Response to Structural Racism*

Background

The Department of Preventive Medicine's (DPM) Working Group on Response to Structural Racism in Research, in collaboration with several members of the Division of Biostatistics, developed these **considerations for best practices** when summarizing and analyzing relevant demographic variables in Human Subjects' Research. These considerations **serve as guidelines and recommendations**. They are not binding in any way. We recognize research studies each have unique needs, and there is no single set of best practices that would apply to all studies.

Considerations for Race/Ethnicity in Reporting and Analyses

Most funding agencies or reporting entities will require summarizations of racial and ethnic distribution (i.e., N/%). In addition, many investigators are interested in quantifying racial/ethnic differences in health outcomes and/or accounting for race/ethnicity as a confounder or effect modifier. Those developing the analysis plan should:

1. Consider the study sample – depending on the source of the data, **availability and format of these data may vary**. This will have implications for summarizations. That is, although it may be of interest to summarize variables by race/ethnicity or to quantify differences in variables across racial/ethnic groups, it may not be analytically possible to do so if the data have not been collected to allow for this or if cell/group counts are too low.
2. **Consider the original goal of collecting these data** in relation to the overarching study goals.
    a. If these data are meant to summarize/describe the study sample only, consider the minimum requirements for the funder, regulatory body, or journal format as a starting point.
    b. If the study aims call for **evaluation of race/ethnicity as a confounder** or **evaluating heterogeneity of effects within racial/ethnic subgroups**, there will be a **compromise** between ensuring the **most granular and precise representation of racial and ethnic identity** versus **degrees of freedom** in statistical model(s).
        i. For example, while there may be interest in evaluating associations within five separate racial or ethnic subgroups, such analyses may not yield stable model estimates. Ideally, all analyses will be pre-specified in the Statistical Analysis Plan (SAP); however, **allow for flexibility in working racial and ethnicity variables into any modeling in the event of small cell counts or violations of basic assumptions**.
        ii. Consider the inference on any coefficients related to race and ethnicity in the context of reporting overall study findings. **Will this practice of adjustment or evaluation of interaction terms involving race truly allow for meaningful and actionable conclusions?**
        iii. **In any case, consider performing analyses both including these variables and excluding**, paying attention to overall differences in inferences. As above, pre-specify these analyses and indicate ahead of time which are the "primary" and which are "secondary", "exploratory", or "sensitivity" analyses.
    c. In any case, **take care to ensure preservation of anonymity to the extent possible**. If there are just a few (e.g., 3 or less individuals [note that specific journals may have specific requirements]) identifying with one racial or ethnic group, consider consolidating the number of categories or adding a footnote explanation for low cell counts for any far-reaching dissemination materials. Note that this may not be possible for funding agencies or regulatory bodies.
3. Options for **parameterization/categorizing groups**:
    a. If each study participant can only fall into one potential category, we suggest **treating the race/ethnicity variable(s) as a factor** in any analytic models. There should be no ordinal nature to the racial or ethnic variables. However, if there are categories with few participants that cause model stability issues, **consider collapsing categories**.

i. **Take care and be sensitive to inferences** that may be drawn as a result of collapsing racial and ethnic groups. For example, is there a priori evidence to suggest the association with the outcome is similar across the racial and ethnic groups that are being collapsed? If not, the true impact of race/ethnicity on the outcome may be misidentified.

ii. **Ensure reporting and dissemination materials clearly indicate the multiple categories** that are combined into one for analyses (e.g., "Black / Multiple Races" may be *one way* to denote two categories that required consolidation into one for statistical reasons). Provide the rationale for this categorization to the extent possible.

iii. Consider an **example study** (based on real data). Participants self-identified with one of the following racial categories: Asian, Black, Multiple Races, or White. The distribution was such that of the 88 participants, 78 (89%) identified as White; 5 (6%) identified with Multiple Races; 4 (4%) identified as Asian; and 1 (1%) participant identified as Black. The study team, after considerable discussions, decided to **collapse the race variable for analyses** and reporting into two categories: "Asian / Black / Multiple Races" (11%) or "White" (89%). As noted above, **meaning of inferences becomes unclear for those interested in any one of the groups represented in the combined group**; however, the alternative was both lack of anonymity and unstable model estimates.

b. If participants had the option to select more than one option for race or ethnicity, we suggest **considering a series of indicator variables (1/0)** for inclusion in model to represent these variables. Using a series of indicator variables rather than mutually exclusive categories allows a participant to self-identify in multiple race/ethnic groups. This may come up if a participant identifies with more than one race group or if a participant identifies as Hispanic/Latino ethnicity and also White race. Consider the guidelines in "a" above in the event of low counts for any one category.

Considerations for Sex as a Biological Variable and Gender

In general, the same recommendations as outlined for race and ethnicity should apply for both sex and gender identity as they pertain to analyses. Of most importance, the statistical team developing and implementing the SAP should determine whether the goals of the study **require collection/reporting/analysis of sex as a biological variable**, **gender identity**, or **both**. Given **NIH often requires considerations for sex as a biological variable**, most analysis plans should plan on summarizing sex (N/%) at minimum, but there will often be plans for evaluation of either confounding or interaction between sex and some other variable. In these instances, refer to points 2 and 3 above. Considerations for gender identity will also mirror those outlined above for race. We suggest the study team follow the guidelines above for variable(s) related to gender identity; however, pay particular attention to the question: **"Will this practice of adjustment or evaluation of interaction terms involving gender identity truly allow for meaningful and actionable conclusions?"**

Potentially Useful References

Kaplan JB, Bennett T. Use of race and ethnicity in biomedical publication. JAMA. 2003 May 28;289(20):2709-16.

Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. New England Journal of Medicine. 2020 Aug 27;383(9):874-82.

***Working Group Members:**

Kiarri Kershaw (Chair); Mercedes Carnethon; Jody Ciolino; Frank Granata; Elizabeth Gray; Mark Huffman; Molly Jones; Monica Rodriguez; Leah Neubauer; Denise Scholtens