

Appendix

A. Multilevel logit model for propensity score subclassification

Propensity score subclassification was used to evaluate and improve covariate balance between the various racial/ethnic subgroups. The propensity scores were estimated using a logit model with the full set of covariates listed in [Table 1](#). The *Stata* 12SE module PSMATCH2 [1,2] was used to examine covariate imbalance pre- and post-matching; see [Supplementary Figure 1](#). Although the use of a matching module such as PSMATCH2 was not necessary for propensity score subclassification, the module includes convenient modeling and diagnostic tools that were executed in *Stata* on the secure server where the NHANES restricted data files were located. The final analysis included estimation of propensity scores using a random effects logit model with the full set of covariates listed in [Table 1](#) as fixed effects, plus a random intercept for county. The random intercept for county was included in the final analysis to account for the county-level heterogeneity in racial and ethnic composition not captured by the covariates in [Table 1](#) and the clustering of the data. In other words, there may be county-level variation in the probability that a participant is non-Hispanic white that is not accounted for by the fixed effects listed in [Table 1](#). Including a random intercept for county addresses this variation, as well as the hierarchical nature of the data where participants are clustered within counties. The Generalized Linear Latent and Mixed Models (GLLAMM) package in *Stata* was used to estimate this final random effects model for the propensity score [3]. Design variables were not included in the propensity score estimation; however, design variables and mobile examination center sample weights were used throughout the remaining analyses to account for the complex sampling design of NHANES. Design variables were not included in the propensity score estimation model for two reasons: first, because variables related to survey selection, participation, and weighting were already included by proxy in the sociodemographic and geographic

covariates; and, second, because the propensity score itself was not intended to generalize to the target population, as it is an in-sample characteristic [4]. As NHANES is a nationally representative sample, propensity-score subclassification was used to avoid discarding observations or modifying the sample weights by using weighting strategies. Five subclasses were used, a convention which has been found to remove 90% or more of the initial covariate bias [5,6].

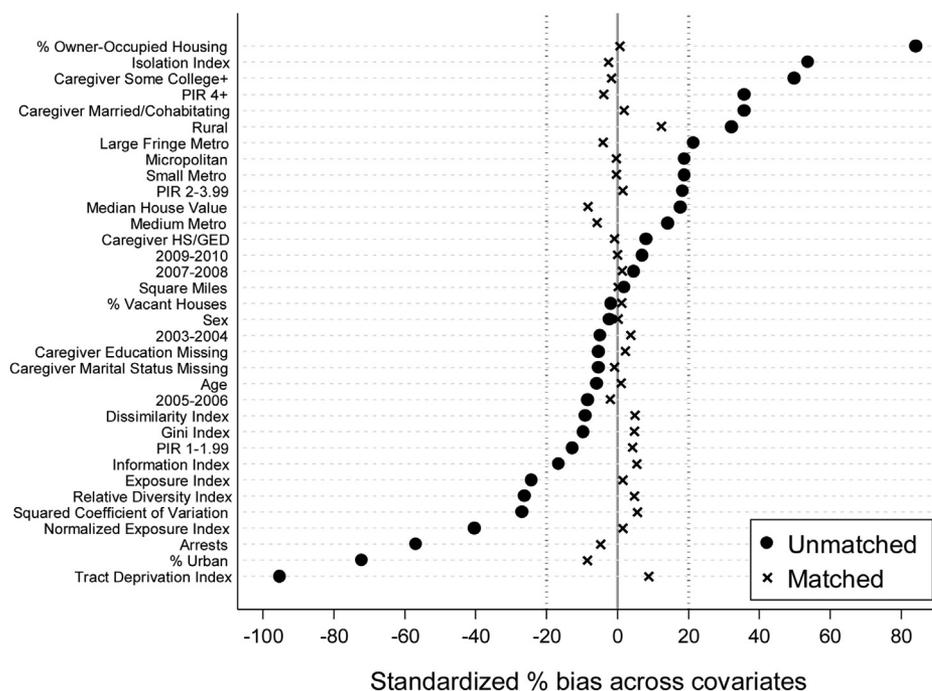
Covariate balance

[Supplementary Figure 1](#) illustrates the standardized bias before propensity score matching, and then after matching (this diagnostic tool was generated by PSMATCH2 using the initial logit model with the full set of covariates listed in [Table 1](#) and excluding random effects). For all included covariates, the remaining bias was less than 20%, and in most cases it was substantially reduced, consistent with a rule of thumb specifying standardized bias be no greater than 25% [5,7]. [Supplementary Figure 2](#) depicts the unweighted sample frequencies of non-Hispanic white and nonwhite children across the range of propensity scores.

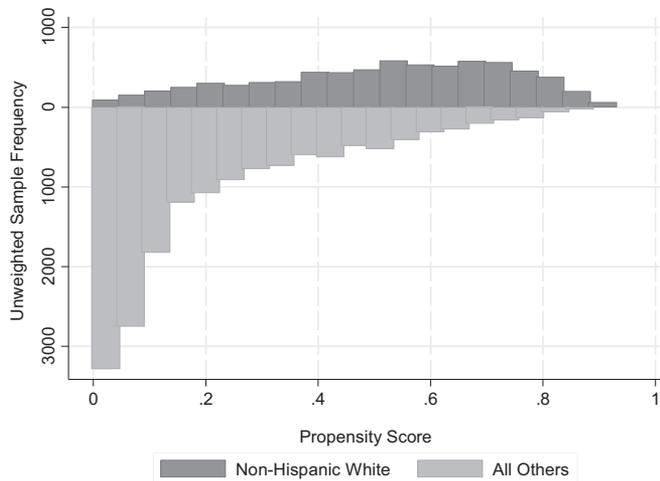
B. Statistical significance of the SRI

To determine statistical significance, the observed Symmetrized Rényi Index (SRI) values were compared with those expected under a null hypothesis of equitability. However, even under the null hypothesis, the sampling distribution of the SRI is unknown [8]. Thus, because of the complex survey design structure of NHANES, a two-stage approach was adopted to assess statistical significance:

- First, the condition of equitability in the outcome variable was simulated by taking the age- and sex-specific mean and standard deviation of the body mass index (BMI) z-scores for the population and simulating normal variates based on those values. As a result, a set of simulated BMI z-scores independent of race/ethnicity and all other covariates was obtained.



Supplementary Figure 1. Standardized bias pre- and post-matching. [‡]For all included covariates, the bias was less than 20% postmatching, and in most cases it was substantially reduced. [‡]Figure generated using the initial propensity score model in PSMATCH2, which did not include the random effect for county.



Supplementary Figure 2. Distribution of propensity scores by race/ethnicity. After propensity-score matching, nonwhite children and adolescents are predominantly grouped in the lower quintile groups, whereas non-Hispanic white children and adolescents fall into the upper quintiles.

- In the second stage, a rescaled bootstrap [8–10] sample (of size 500) allowed for the design-based estimation of the 99th percentile of the sample SRI for this set of simulated BMI z-scores, which enabled testing of the null hypothesis that $SRI = 0$ against the alternative that SRI greater than 0 at the 1% significance level. To improve the estimation of the null 99th percentile, which is the basis for the hypothesis rejection rule, the aforementioned BMI z-score simulations were repeated 100 times and, to guard against extreme values, a 10% trimmed mean was constructed as a robust estimate of the null 99th percentile. This type of averaging over bootstrap samples (or “bagging”) is known to improve the performance of quantile estimators [11–14]. This approach is particularly suited to the analysis, here, because of the two distinct sources of variability: (i) variability in the outcome variable (BMI z-score) under the

null hypothesis of equitability and (ii) the design-based sample variability (encoded in the survey weights, strata, and primary sampling units). Design-based estimation of the SRI and of the 99th percentile of its distribution under the null was conducted in R using the “survey” package and code developed by the second author [15–18].

Supplemental References

- [1] StataCorp. Stata Statistical Software. Release. 12 ed. College Station, TX: StataCorp LP; 2011.
- [2] Leuven E, Sianesi B. PSMATCH2: Stata module to perform full Mahalanobis and propensity score matching, common support graphing, and covariate imbalance testing. 2012.
- [3] Rabe-Hesketh S, Skrondal A, Pickles A. GLLAMM manual. Berkeley, CA: 2004.
- [4] Dugoff EH, Schuler M, Stuart EA. Generalizing observational study results: applying propensity score methods to complex surveys. *Health Serv Res* 2014;49(1):284–303.
- [5] Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat Sci* 2010;25(1):1–21.
- [6] Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984;79(387):516–24.
- [7] Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods* 2010;15(3):234–49.
- [8] Talih MA. Reference-invariant health disparity index based on Rényi divergence. *Ann Appl Stat* 2013;7(2):1217–43.
- [9] Cheng NF, Han PZ, Gansky SA. Methods and software for estimating health disparities: the case of children’s oral health. *Am J Epidemiol* 2008;168(8):906–14.
- [10] Rao JNK, Wu CFJ. Resampling inference with complex survey data. *J Am Stat Assoc* 1988;83(401):231–41.
- [11] Buhlmann P, Yu B. Analyzing bagging. *Ann Stat* 2002;30(4):927–61.
- [12] Buja A, Stuetzle W. Observations on bagging. *Stat Sinica* 2006;16(2):323–51.
- [13] Knight K, Bassett GW. Second order improvements of sample quantiles using subsamples. Tech Rep [Internet]. 2005. Available from: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.9.1424>.
- [14] Wang JC, Opsomer JD, Wang H. Bagging non-differentiable estimators in complex surveys. Tech Rep [Internet]. 2012. Available from: <http://www.stat.colostate.edu/~jopsomer/papers/bagging.pdf>.
- [15] Lumley T. Analysis of complex survey samples. *J Stat Softw* 2004;9:1–9.
- [16] Lumley T. “Survey”: analysis of complex survey samples. In: Lumley T, editor. R package version 3.26 ed 2011.
- [17] R Development Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
- [18] Talih M. Supplement to “A reference-invariant health disparity index based on Rényi divergence”—R syntax and output files. *Ann Appl Stat* 2013;7(2):1217–43.