HSR Health Services Research

DOI: 10.1111/1475-6773.12295
METHODS BRIEF

Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data

Robert W. Grundmeier, Lihai Song, Mark J. Ramos, Alexander G. Fiks, Marc N. Elliott, Allen Fremont, Wilson Pace, Richard C. Wasserman, and Russell Localio

Objective. To assess the utility of imputing race/ethnicity using U.S. Census race/ethnicity, residential address, and surname information compared to standard missing data methods in a pediatric cohort.

Data Sources/Study Setting. Electronic health record data from 30 pediatric practices with known race/ethnicity.

Study Design. In a simulation experiment, we constructed dichotomous and continuous outcomes with pre-specified associations with known race/ethnicity. Bias was introduced by nonrandomly setting race/ethnicity to missing. We compared typical methods for handling missing race/ethnicity (multiple imputation alone with clinical factors, complete case analysis, indicator variables) to multiple imputation incorporating surname and address information.

Principal Findings. Imputation using U.S. Census information reduced bias for both continuous and dichotomous outcomes.

Conclusions. The new method reduces bias when race/ethnicity is partially, nonrandomly missing.

Key Words. Multiple imputation, U.S. Census location and surname data, race and ethnicity, health disparities

Comparative effectiveness research using electronic health record (EHR) data promises timely understanding of health disparities (Olsen, Aisner, and McGinnis 2007; Fiks et al. 2012) but requires accurate racial/ethnic data, which are often completely or partially missing in EHR data (Bierman et al. 2002; Kressin et al. 2003; Hasnain-Wynia, Pierce, and Pittman 2004; Fremont et al. 2005; Bilheimer and Sisk 2008; Smith et al. 2010; Wynia, Ivey, and Hasnain-Wynia 2010; Bergdall et al. 2012).

Novel analytic methods for missing EHR race/ethnicity data are needed, as many current methods have shortcomings. Complete case analysis confined to observations without missing values produces biased estimates, except when data are missing completely at random (MCAR)—unassociated with any outcomes or covariates—or when the chosen statistical model assumes only missingness at random (Molenberghs and Kenward 2007; National Research Council (United States), Panel on Handling Missing Data in Clinical Trials, National Research Council (United States), and Committee on National Statistics 2010; Graham 2012). The missing indicator method—using a missing category indicator for missing variables such as race/ethnicity—also produces biased estimates and (Greenland and Finkle 1995; Knol et al. 2010). Mean imputation can understate variance and underestimate *p*-values. Multiple imputation, though effective when used appropriately, is inefficient compared to an analysis based on all data if no values were missing, especially when a large fraction of the data is missing or key predictors of missing values are absent (Rubin 1987).

To address problems of missing race/ethnicity in disparities research, RAND researchers created the Bayesian Improved Surname Geocoding (BISG) method to impute race/ethnicity among adult patients using U.S. Census geospatial and Census surname data (Fiscella and Fremont 2006; Elliott et al. 2008, 2009). Subsequently, a categorized version of the 2009 algorithm was validated for classifying patients into racial/ethnic groups, with favorable results across the age spectrum (Adjaye-Gbewonyo et al. 2014), and a couples version was validated in a study of marriage licenses (Elliott et al. 2013). The Centers for Medicare & Medicaid Services and health plans have successfully used the BISG to better understand racial/ethnic disparities (Derose et al. 2013; Martino et al. 2013).

Here, we used the BISG's probability outputs in a pediatric cohort to enhance generally accepted methods of imputing missing race/ethnicity data.

Address correspondence to Robert W. Grundmeier, M.D., The Children's Hospital of Philadelphia, 3535 Market Street, Philadelphia, PA 19104; e-mail: grundmeier@email.chop.edu. Lihai Song, M.S., Mark J. Ramos B.S., and Alexander G. Fiks, M.D., M.S.C.E., are also with The Children's Hospital of Philadelphia, Philadelphia, PA. Marc N. Elliott, Ph.D., and Allen Fremont, M.D., are with the RAND Corporation, Santa Monica, CA. Wilson Pace, M.D., is with the DART-Net Institute, University of Colorado Denver, Aurora, CO. Richard C. Wasserman, M.D., M.P.H., is with the University of Vermont, Burlington, VT. Russell Localio, Ph.D., is with the Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA.

METHODS

Using a retrospective cohort design, we added BISG probability outputs to pediatric data to evaluate whether these probabilities improve existing methods of imputing missing race/ethnicity. We extracted EHR data for children with known race/ethnicity and experimentally assigned race/ethnicity to "missing" for a sample of these children to measure the performance of different strategies for estimating missing race/ethnicity. We included children who received preventive health care at 30 practices in Pennsylvania and New Jersey between July 1, 2009, and June 30, 2010, before their 19th birthday. The 30 practices routinely recorded parent-reported race/ethnicity during registration. Children were identified by preventive health care evaluation and management codes in the EHR (CPT codes 99381-99385, 99391-99395).

Beginning with 220,090 children, we excluded 26,007 children whose address could not be accurately geocoded, 1,608 for whom U.S. Census race/ethnicity data were not available for their census block group, 19,447 with unknown or low-frequency race/ethnicity values (e.g., Native American, multiracial), and 2,857 children without insurance, leaving 170,171 children.

Study variables included demographic and clinical characteristics possibly associated with race/ethnicity. In addition to parent-reported race and ethnicity, these variables were child surname, home address, gender, age and insurance type, presence of asthma, and presence of attention-deficit hyperactivity disorder (ADHD) as identified by ICD-9 codes specific to these conditions on the problem list (codes beginning with the digits 493 and 314, respectively).

Bayesian Improved Surname Geocoding Method

The BISG method estimates a posterior probability of membership in each of six racial/ethnic categories—(1) Hispanic/Latino; or non-Hispanic (2) black/African American; (3) white; (4) Asian/Pacific Islander; (5) Native American/Alaska Native; or (6) multiracial—using the patient's surname and the U.S. Census block group of their home address. The BISG algorithm in this instance used 2000 U.S. Census data about the race/ethnicity of each block group and the probability of each surname belonging to each of the six racial/ethnic categories above. Race/ethnicity information is available only for surnames that occur at least 100 times in the U.S. population (Word et al. 2008); for rare surnames, the algorithm uses aggregate race/ethnicity information for

all U.S. Census respondents whose surnames occur fewer than 100 times (Elliott et al. 2009).

Geocoding Address to Census Block Group

Patient address was geocoded to the corresponding latitude, longitude, and 2000 census block group using the *Geocoder/U.S.* v2.0 software package (Erle n.d.), which scored the confidence of the match with a 0 to 1 score. Scores less than 1 indicated the need for corrections (e.g., misspelled street names). 194,083 (88 percent) of the children in our cohort were geocoded with a score of 0.8 or higher. To ensure accurate geocoding, we manually validated a random sample of 50 addresses with scores in the 0.8 to 0.9 range and confirmed that all were geocoded correctly.

Statistical Methods

As a first comparison, we used logistic regression to model the true racial/ethnic category as a function of the imputed value. This model produced an area under a receiver operating characteristic curve (AROC) as a measure of the predicted value's ability to discriminate between the given racial/ethnic category and other categories.

Iterative Simulations. We iteratively performed 200 simulations with the following steps: (1) generated hypothetical outcomes, both continuous and binary; (2) set known race/ethnicity to missing; (3) applied four strategies for analyzing data in the presence of missing values; and (4) used regression models to estimate the association of the outcomes with the racial/ethnic groups. We compared the true values for the hypothetical outcomes with the median values and interquartile ranges for each analytic method: complete case analysis, missing value indicator method, imputation using patient information, and imputation with BISG probabilities added to patient information. The simulation analysis was performed on a 10 percent sample of children (n = 17,017) for computational efficiency.

Generating Outcome Variables. To test our ability to derive unbiased estimates for the association of race/ethnicity with outcome, we constructed two outcome variables strongly associated with race/ethnicity: one continuous and one

dichotomous (equations 1 and 2). The continuous outcome represented a normally distributed score similar to an ADHD symptom score (Wolraich et al. 1998), with mean of 100 and standard deviation of 10. The dichotomous outcome represented an endpoint such as pediatric asthma hospitalization, often strongly associated with black/African American race/ethnicity. We constructed these outcomes to have plausible associations with race/ethnicity and disease status that could result in bias if the data were not missing randomly.

Race/Ethnicity Missing Not at Random. Using the complete dataset with the added outcomes noted above, we then set race/ethnicity to missing for a fraction of the observations. For the continuous outcome, we used an equation that increased the chances of being missing for black children with an outcome score that was less than 90, and for Hispanic children whose outcome scores exceeded 105. This equation correlated the probability of missingness with true race/ethnicity and also with the value of the outcome, resulting in a nonignorable (or Missing Not At Random) missing data mechanism. Similarly, for the binary outcome simulations, we increased the chance of missingness for black children when the outcome was 0 and for Hispanic children where the outcome was 1 (Table 1).

Missing Data Methods. We used four strategies to handle missing race/ethnicity: (1) complete case analysis, using only children with nonmissing data; (2) missing indicator method; (3) traditional multiple imputation without BISG enhancement (Schafer 1999); and (4) traditional imputation enhanced by the BISG method.

For traditional multiple imputation, we used variables that might be predictive of race/ethnicity. We used the method of chained equations with race/ethnicity as a multinomial outcome in the regression model to impute race/ethnicity from the following patient-level covariates: gender, age in years, insurance type, diagnosis of asthma, and diagnosis of ADHD (Royston 2004; Van Buuren 2007; White, Royston, and Wood 2011). In addition, because the location of care in a highly diverse major city and its suburbs might be highly predictive of race/ethnicity, we included as a series of 29 indicator variables the primary care practice site. Finally, as recommended (Little 1992), we used the dichotomous and continuous outcome scores as predictors in the imputation model. In the enhanced imputation, we used these same predictor variables—including care location—and the probability of membership in the

Table 1: Methods for Generating Hypothetical Outcomes (Equations 1 and 2) and Then for Setting Known Race Ethnicity Categories to Missing (Equations 3 and 4) for Purposes of Testing the Performance of Missing Data Imputation Methods

Equation Name	Equation Definition
Continuous and dichotomous outcomes	
Equation 1: Continuous outcome	$\text{Score}_c = 100 - 10 * RE_{\text{black}} + 10 * RE_{\text{hispanic}} - 5 * Dx_{\text{asthma}} + e$ (1)
Equation 2: Dichotomous outcome	$\text{log-odds}(\text{E}(\text{Score})) = \mu_{\text{outcome}} = \text{logit}(0.2) - 0.693 * RE_{\text{black}} + 0.693 * RE_{\text{hispanic}} + 0.289 * Dx_{\text{asthma}}$ (2)
Setting race and ethnicity to missing to introduce bias	
Equation 3: Bias for continuous outcome	$\mu_{\text{missing}} = \text{logit}(0.2) + 2.25 * RE_{\text{black}} * (\text{Score}_c < 90) + 1.25 * RE_{\text{Hispanic}} * (\text{Score}_c > 105)$ (3)
Equation 4: Bias for dichotomous outcome	$\mu_{missing} = \text{logit}(0.2) + 2.25 * RE_{\text{black}} * (\text{Score} = 0) + 1.25 * RE_{\text{Hispanic}} * (\text{Score} = 1)$ (4)

Variable Name	Definition
Score	Outcome score (c = continuous)
RE	Race/ethnicity
Dx	Indicator variable for asthma or ADHD
e	Error term, normally distributed, mean = 0, standard deviation = 20
μ_{outcome}	Mean of a binary score drawn randomly from a binomial distribution
μ_{missing}	Log odds of the predicted probability of race being missing

six race/ethnicity categories calculated with the U.S. Census geospatial and surname data (the BISG method's output). For each of the 200 iterations in this simulation, we repeated the imputation to construct 10 datasets for both the traditional imputation and the enhanced imputation.

Estimating the Association of Race and Ethnicity with Outcome

The associations between demographic characteristics (race/ethnicity, gender, age, and primary care site) and the outcome variables were calculated using multivariable linear regression for the continuous outcome and multivariable logistic regression for the dichotomous outcome. We compared the associations observed in each of the experimental datasets with the true associations observed in the gold standard dataset.

Linear and logistic regressions were performed using *Stata v 13.0* (2008; Stata Corp, College Station TX, USA). The Institutional Review Board at The Children's Hospital of Philadelphia approved the study and waived the requirement for consent from individual children/families.

RESULTS

Demographic and clinical characteristics for the study cohort appear in Table 2.

Performance of BISG-Supplemented Imputation

The surname for 171,093 (88 percent) of the 194,083 geocoded patients exactly matched entries in the U.S. Census surname list, similar to the population rates of common surnames (Elliott et al. 2009). As expected from the construction of the Census surname list (Word et al. 2008), the unmatched surnames included hyphenated names, misspellings of common names, and uncommon names. All unmatched surnames were included as "rare surnames" (see BISG methods section).

Ability of BISG-Supplemented Imputation to Identify True Race/Ethnicity

For the major race/ethnicity categories, the census-based probabilities had good qualities as a test for parent-reported race and ethnicity. The AROC was high for all race and ethnicities prevalent in our cohort (Table 3).

Table 2: Characteristics of the Study Cohort

Group	*N*	*%*
Hispanic or Latino ethnicity	8,706	5.1
Non-Hispanic		
White	101,377	59.6
Black or African American	55,286	32.5
Asian or Pacific Islander	4,802	2.8
Gender		
Female	83,700	49.2
Male	86,471	50.8
Age		
Birth to 4 years	54,987	32.3
5 to 11 years	65,950	38.8
12 years and up	49,234	28.9
Disease		
Asthma	22,216	13.1
Attention-deficit disorder	7,462	4.4

Estimated Racial/Ethnic Proportions in the Cohort

By design, the distribution of race/ethnicity among cases with nonmissing data differed substantially from the true proportions. In all categories, the addition of the BISG probabilities to formal imputation models slightly improved the estimates for the proportion of children in each race/ethnicity category. In most categories the addition of care location also resulted in slight improvements (Table 4).

Performance of Imputation Methods

Median values from simulations for the association of race/ethnicity and outcome appear in Table 5. As expected by the design, the regression estimates obtained using all data in the gold-standard dataset matched the stipulated values for both the continuous and dichotomous outcomes. Two methods for handling missing data—complete case analysis in which only data with nonmissing values are used and the missing indicator variable method—produced highly biased results. With the binary outcome for these two methods, the direction for the odds ratio was reversed for the black subjects (an odds ratio greater than 1 was estimated when the true odds ratio was 0.5).

Standard multiple imputation reduced the bias somewhat, especially for the estimates of both the continuous and dichotomous outcome associated with black race. However, in the case of the dichotomous outcome,

Table 3: Area under the Receiver Operating Characteristic (AROC) Curve as a Measure of the Ability of the BISG Measures of Race/Ethnicity to Discriminate among the True Race and Ethnicity Groups

*Race or Ethnicity**	*N*	*ROC Area*[†]	*% Correctly Identified*[‡]
Hispanic or Latino	8,706	0.88	50.2
White	101,377	0.95	97.2
Black or African American	55,286	0.97	75.8
Asian or Pacific Islander	4,802	0.91	62.6
Total	170,171	0.95[§]	

*American Indian/Alaska Natives and multiracial categories were excluded.
[†]ROC area compares the performance of the BISG probability membership in a particular race or ethnic group as a continuous score (as in a diagnostic test score) compared to the true dichotomous value for the child's membership in that group (i.e., 1 if the child is a member of that race/ethnic group, versus 0 if the child is not).
[‡]Reports the percent of children in each race/ethnic group for which the BISG algorithm assigned the highest probability to the correct group. This metric corresponds to the less accurate dichotimization of BISG probabilities (Adjaye-Gbewonyo et al. 2014).
[§]Population-weighted average of area under the receiver operating characteristic curve. This metric corresponds to the more accurate direct uses of BISG probabilities (Elliott et al. 2009).

Table 4: Estimated Proportion of Races by Method of Imputation in the Sample Cohort (Excludes Multiracial Designations)

	Before Imputation		*Traditional Imputation*		*BISG-Enhanced Imputation*	
Group[†]	*Parent-Report (%)*[‡]	*Exclude Missing (%)*[§]	*No Site (%)*	*With Site (%)*	*No Site (%)*	*With Site (%)*
Hispanic	5.1	4.9	4.9	4.8	4.9	5.0
White	59.6	67.6	65.2	62.8	61.9	61.4
Black[¶]	32.5	24.5	26.8	29.3	30.4	30.7
Asian	2.8	4.9	3.1	3.1	2.8	2.9

Notes: $N = 17{,}105$; 10% random sample of the cohort. These results are not based on simulations.
[†]American Indian/Alaska Natives and multiracial categories were dropped.
[‡]Parent-report reflects the actual report of race/ethnicity in the study practices prior to setting data to missing.
[§]Data were assigned to missing in a nonrandom fashion and then imputed using standard and census enhanced imputation methods.
[¶]Black race was significantly under-represented after data were set to missing. Adding additional predictors (primary care practice site and BISG probabilities) to multiple imputation partially recovered the true proportions.

the odds ratio estimate using multiple imputation was only slightly less than 1 for black race/ethnicity (true value 0.5). Bias fell substantially, although it did not disappear, for BISG-supplemented-multiple imputation.

Table 5: Bias of Different Methods of Estimating Racial/Ethnic Differences: True Values and Estimates by Method for Continuous and Binary Outcomes

Race/Ethnicity	*True Value*	*All Data**	*Complete Case Analysis*	*Indicator Variable Method*	*Multiple Imputation*	*BISG-Enhanced Imputation*
Continuous outcome[†]						
Black	−10	−10.0 [−10.3, −9.71]	−2.59 [−2.89, −2.23]	−0.96 [−1.26, −0.70]	−2.96 [−3.41, −2.49]	−8.05 [−8.31, −7.74]
Hispanic	+10	9.97 [9.52, 10.6]	6.88 [6.09, 7.47]	8.21 [7.38, 8.79]	7.62 [6.86, 8.29]	8.30 [7.61, 8.88]
Binary outcome[‡]						
Black[§]	0.5	0.50 [0.48, 0.52]	1.35 [1.29, 1.41]	1.54 [1.48, 1.60]	1.30 [1.22, 1.39]	0.65 [0.62, 0.68]
Hispanic	2.0	1.99 [1.89, 2.11]	1.33 [1.26,1.42]	1.48 [1.40, 1.58]	1.34 [1.27, 1.45]	1.55 [1.45, 1.64]

*All data in the gold-standard dataset were included to confirm that in the ideal situation with no data missing the regression will closely estimate the true value.
[†]Median values and interquartile ranges of estimated coefficients from linear regression in 200 simulations.
[‡]Median values and interquartile ranges of estimated odds ratios from logistic regression in 200 simulations.
[§]In the case of the binary outcome bias was extreme for black race (the true odds ratio was 0.5, but the estimated odds ratio was more than 1 for complete case analysis, indicator variable method, and multiple imputation when only patient data was included). Adding BISG probabilities to multiple-imputation estimated coefficients (for the continuous outcome) and odds ratios (for the binary outcome) that were closer to the true value than the other methods.

In the case of the dichotomous outcome, only the BISG-supplemented approach correctly estimated the odds ratio associated with black race/ethnicity to be less than 1.

As a final check, we assessed whether the use of the BISG probabilities introduced bias in multiple imputation when race/ethnicity data were missing in an unbiased fashion. We found that no bias was introduced by the use of BISG data in this scenario (data not shown).

DISCUSSION

One promise of "big data" is that large datasets derived from EHR and administrative sources will expedite understanding of health disparities. Unfortunately, large clinical data research networks that aggregate data from multiple health systems, such as those funded by the Patient-Centered Outcomes Research Institute, will encounter randomly and nonrandomly missing race/ethnicity data for the foreseeable future.

We investigated whether the addition of the new information derived from the BISG surname/address method to traditional methods of multiple imputation, improves the accuracy of estimates in the presence of partially missing data on race and ethnicity in a large pediatric EHR dataset. We deliberately used conventional missing data methods because the software for those methods is increasingly available and the output provides discrete categories rather than probabilities, allowing reporting both from regression analyses and contingency tables. A case of extreme bias with a calculated odds ratio in the wrong direction for a dichotomous outcome was almost completely corrected using the BISG-enhanced imputation approach. In this case, the BISG-enhanced imputation odds (0.65) were similar to the true odds (0.5). In our cohort, the enhanced imputation model improved estimates and reduced bias. An additional advantage of formal imputation of missing data, as contrasted with use of only complete data, lies in the increased statistical power of the larger usable samples.

Limitations

Although the 30 primary care sites in this study care for a racially and ethnically diverse group of children, the sample was limited to a single health system in two states. The data for this study were derived from EHRs captured as part of routine care and likely include misclassification errors in all variables.

We excluded multiracial children from the dataset because there was no reliable way to enter this classification in our EHR. Multiracial children were typically described as having "other" or "unknown" race/ethnicity at the time of this study. Also, as less than 1 percent of the children in our cohort were American Indian, we were unable to evaluate performance for this subgroup. Insurance type was our only marker for socioeconomic status. More precise markers such as household income should be included in multiple imputation when available.

Race/ethnicity probabilities from the BISG algorithm do not include estimates of the uncertainty of those probabilities. However, uncertainty in BISG probabilities has been shown to have negligible effects on estimation under reasonable assumptions (McCaffrey and Elliott 2008). Finally, with increasing numbers of racially blended families, the surnames of children over time might have decreasing value in predicting racial/ethnic backgrounds.

Conclusions

When used with standard statistical tools for proper imputation of missing values, the additional data from an algorithm for identifying race/ethnicity using U.S. Census data on surnames and home address lead to reduced bias in typical analyses of associations between race/ethnicity and health outcomes. Our results suggest that, even in pediatric settings, this algorithm should be preferred over standard methods for addressing missing race/ethnicity data.

Acknowledgments

Joint Acknowledgment/Disclosure Statement: This research was supported by grant funding to the American Academy of Pediatrics from the Maternal & Child Health Bureau (UB5MC20286).

Disclosures: None.

Disclaimers: None.

References

Adjaye-Gbewonyo, D., R. A. Bednarczyk, R. L. Davis, and S. B. Omer. 2014. "Using the Bayesian Improved Surname Geocoding Method (BISG) to Create a

Working Classification of Race and Ethnicity in a Diverse Managed Care Population: A Validation Study." *Health Services Research* 49 (1): 268–83. doi:10.1111/1475-6773.12089.

Bergdall, A., S. Asche, N. Schneider, T. Kerby, K. Margolis, J. Sperl-Hillen, J. Sekenski, R. Pritchard, M. Maciosek, and P. O'Connor. 2012. "CB3-01: Comparison of Ethnicity and Race Categorization in Electronic Medical Records and by Self-Report." *Clinical Medicine & Research* 10 (3): 172. doi:10.3121/cmr.2012.1100.cb3-01.

Bierman, A. S., N. Lurie, K. S. Collins, and J. M. Eisenberg. 2002. "Addressing Racial and Ethnic Barriers to Effective Health Care: The Need for Better Data." *Health Affairs (Project Hope)* 21 (3): 91–102.

Bilheimer, L. T., and J. E. Sisk. 2008. "Collecting Adequate Data on Racial and Ethnic Disparities in Health: The Challenges Continue." *Health Affairs (Project Hope)* 27 (2): 383–91. doi:10.1377/hlthaff.27.2.383.

Derose, S. F., R. Contreras, K. J. Coleman, C. Koebnick, and S. J. Jacobsen. 2013. "Race and Ethnicity Data Quality and Imputation Using U.S. Census Data in an Integrated Health System: The Kaiser Permanente Southern California Experience." *Medical Care Research and Review* 70 (3): 330–45. doi:10.1177/1077558712466293.

Elliott, M. N., A. Fremont, P. A. Morrison, P. Pantoja, and N. Lurie. 2008. "A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity." *Health Services Research* 43 (5 Pt 1): 1722–36. doi:10.1111/j.1475-6773.2008.00854.x.

Elliott, M. N., P. A. Morrison, A. Fremont, D. F. McCaffrey, P. Pantoja, and N. Lurie. 2009. "Using the Census Bureau's Surname List to Improve Estimates of Race/Ethnicity and Associated Disparities." *Health Services and Outcomes Research Methodology* 9 (2): 69–83. doi:10.1007/s10742-009-0047-1.

Elliott, M. N., K. Becker, M. K. Beckett, K. Hambarsoomian, P. Pantoja, and B. Karney. 2013. "Using Indirect Estimates Based on Name and Census Tract to Improve the Efficiency of Sampling Matched Ethnic Couples from Marriage License Data." *Public Opinion Quarterly* 77 (1): 375–84. doi:10.1093/poq/nft007.

Erle, S. n.d. "Geocoder: U.S." [accessed on May 24, 2013]. Available at https://github.com/geocommons/geocoder

Fiks, A. G., R. W. Grundmeier, B. Margolis, L. M. Bell, J. Steffes, J. Massey, and R. C. Wasserman. 2012. "Comparative Effectiveness Research Using the Electronic Medical Record: An Emerging Area of Investigation in Pediatric Primary Care." *Journal of Pediatrics* 160 (5): 719–24. doi:10.1016/j.jpeds.2012.01.039.

Fiscella, K., and A. M. Fremont. 2006. "Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity." *Health Services Research* 41 (4 Pt 1): 1482–500. doi:10.1111/j.1475-6773.2006.00551.x.

Fremont, A. M., A. Bierman, S. L. Wickstrom, C. E. Bird, M. Shah, J. J. Escarce, T. Horstman, and T. Rector. 2005. "Use of Geocoding in Managed Care Settings to Identify Quality Disparities." *Health Affairs (Project Hope)* 24 (2): 516–26. doi:10.1377/hlthaff.24.2.516.

Graham, J. W. 2012. *Missing Data Analysis and Design.* New York: Springer.

Greenland, S., and W. D. Finkle. 1995. "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses." *American Journal of Epidemiology* 142 (12): 1255–64.

Hasnain-Wynia, R., D. Pierce, and M. A. Pittman. 2004. *Who, When, and How: The Current State of Race, Ethnicity, and Primary Language Data Collection in Hospitals.* The Commonwealth Fund #726 [accessed February 22, 2015]. Available at: http://www.commonwealthfund.org/~/media/files/publications/fund-report/2004/may/who--when--and-how--the-current-state-of-race--ethnicity--and-primary-language-data-collection-in-ho/hasnain-wynia_whowhenhow_726-pdf.pdf

Knol, M. J., K. J. M. Janssen, A. R. T. Donders, A. C. G. Egberts, E. R. Heerdink, D. E. Grobbee, K. G. M. Moons, and M. I. Geerlings. 2010. "Unpredictable Bias When Using the Missing Indicator Method or Complete Case Analysis for Missing Confounder Values: An Empirical Example." *Journal of Clinical Epidemiology* 63 (7): 728–36. doi:10.1016/j.jclinepi.2009.08.028.

Kressin, N. R., B.-H. Chang, A. Hendricks, and L. E. Kazis. 2003. "Agreement between Administrative Data and Patients' Self-Reports of Race/Ethnicity." *American Journal of Public Health* 93 (10): 1734–9.

Little, R. J. A. 1992. "Regression with Missing X's: A Review." *Journal of the American Statistical Association* 87 (420): 1227–37.

Martino, S. C., R. M. Weinick, D. E. Kanouse, J. A. Brown, A. M. Haviland, E. Goldstein, J. L. Adams, K. Hambarsoomian, D. J. Klein, and M. N. Elliott. 2013. "Reporting CAHPS and HEDIS Data by Race/Ethnicity for Medicare Beneficiaries." *Health Services Research* 48 (2 Pt 1): 417–34. doi:10.1111/j.1475-6773.2012.01452.x.

McCaffrey, D. F., and M. N. Elliott. 2008. "Power of Tests for a Dichotomous Independent Variable Measured with Error." *Health Services Research* 43 (3): 1085–101. doi:10.1111/j.1475-6773.2007.00810.x.

Molenberghs, G., and M. G. Kenward. 2007. *Missing Data in Clinical Studies, Chichester.* Hoboken, NJ: Wiley.

National Research Council (U.S.), Panel on Handling Missing Data in Clinical Trials, National Research Council (U.S.), and Committee on National Statistics. 2010. *The Prevention and Treatment of Missing Data in Clinical Trials.* Washington, DC: National Academies Press.

Olsen, L., D. Aisner, J. M. McGinnis, and Institute of Medicine (U.S.), and Roundtable on Evidence-Based Medicine. 2007. *The Learning Healthcare System Workshop Summary.* Washington, DC: National Academies Press.

Royston, P. 2004. "Multiple Imputation of Missing Values." *Stata Journal* 4 (3): 227–41.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys.* New York: Wiley.

Schafer, J. L. 1999. "Multiple Imputation: A Primer." *Statistical Methods in Medical Research* 8 (1): 3–15.

Smith, N., R. L. Iyer, A. Langer-Gould, D. T. Getahun, D. Strickland, S. J. Jacobsen, W. Chen, S. F. Derose, and C. Koebnick. 2010. "Health Plan Administrative Records versus Birth Certificate Records: Quality of Race and Ethnicity Information in Children." *BMC Health Services Research* 10: 316. doi:10.1186/1472-6963-10-316.

Van Buuren, S. 2007. "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification." *Statistical Methods in Medical Research* 16 (3): 219–42. doi:10.1177/0962280206074463.

White, I. R., P. Royston, and A. M. Wood. 2011. "Multiple Imputation Using Chained Equations: Issues and Guidance for Practice." *Statistics in Medicine* 30 (4): 377–99. doi:10.1002/sim.4067.

Wolraich, M. L., I. D. Feurer, J. N. Hannah, A. Baumgaertel, and T. Y. Pinnock. 1998. "Obtaining Systematic Teacher Reports of Disruptive Behavior Disorders Utilizing DSM-IV." *Journal of Abnormal Child Psychology* 26 (2): 141–52.

Word, D. L., C. D. Coleman, R. Nunziata, and R. Kominski. 2008. "Demographic Aspects of Surnames from Census 2000" [accessed on January 30, 2014]. Available at http://www.census.gov/genealogy/www/surnames.pdf

Wynia, M. K., S. L. Ivey, and R. Hasnain-Wynia. 2010. "Collection of Data on Patients' Race and Ethnic Group by Physician Practices." *New England Journal of Medicine* 362 (9): 846–50. doi:10.1056/NEJMsb0910799.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article:

Appendix SA1: Author Matrix.