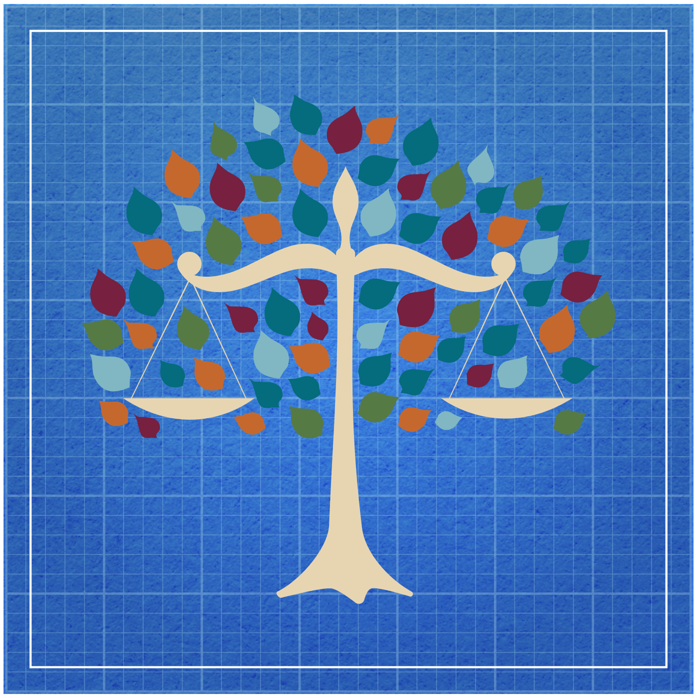
OCTOBER 2022



WHITE PAPER

Inferring Race and Ethnicity Demographics from U.S. Census Data:

Testing the Feasibility for Use in State Court Disparity Analyses

By Miriam Hamilton

NCSC
National Center for State Courts

Introduction

To enable leaders to make data-driven decisions that improve racial equity in the courts, courts must have access to valid race and ethnicity data and regularly use those data to monitor for disparities in court outcomes. Yet, court researchers, like researchers in numerous fields, face severe obstacles due to the lack of robust data on race and ethnicity. Self-reported data is considered the gold standard for collecting demographic information.[2] However, the collection of these data can be challenging or even impossible in some instances. Self-reported data are limited by nonresponses, as people often refuse to provide these data.[3] Additionally, courts often lack the staff time needed to collect the information. Those courts that collect race and ethnicity data frequently rely on observation based on the physical characteristics of an individual.[4] Yet, when jurisdictions rely on observation, the unstandardized reporting becomes unreliable, creating validity concerns for sharing the data. Additional barriers include the lack of interoperable systems and other technical issues, as well as various concerns about the data being misused or misinterpreted.[5]

Improving data collection is crucial for courts as it is in other fields. A variety of efforts to improve justice system data, including court data, is underway nationally. One example is the National Open Court Data Standards (NODS), initiated by the Conference of State Court Administrators (COSCA) and developed by the National Center for State Courts (NCSC) to support court data creation, sharing, and integration.[6] While efforts to improve court data continue in earnest, the need to identify and address racial and ethnic disparities in court outcomes remains urgent. Researchers are looking for ways to address missing or poor-quality race and ethnicity data so disparity analyses may be conducted. Thus, techniques to impute demographic data, including race and ethnicity, have been developed and refined to enable disparity monitoring and analysis.

Researchers employing methods to impute race/ethnicity data must understand the complex methodological issues involved in measuring race/ethnicity, as these will inevitably also restrict decisions about the imputation of such data and the ways the validity of the method should be measured. Adding to

[1] Approved for publication by the Blueprint for Racial Justice Steering Committee on October 14, 2022.

[2] Kathryn Genthon & Diane Robinson, *Collecting Race & Ethnicity Data*, National Center for State Courts (2022), https://www.court-statistics.org/__data/assets/pdf_file/0036/69678/Race_Ethnicity_Data_Collection_4.pdf; Kevin Fiscella & Allen M. Fremont, *Use of Geocoding and Surname Analysis to Estimate Race and Ethnicity*, 41 Health Serv. Res.1482, 1496 (2006).

[3] Fiscella and Fremont, supra note 2 at 1483; Marc N. Elliott et al., *Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities*, 9 Health Serv. Outcomes Res. Methodol. 69, 70 (2009).

[4] Kathryn Genthon & Diane Robinson, *supra* note 2.

[5] *Id.*

[6] NODS resources are available at www.ncsc.org/nods. Another example is the Justice Counts Initiative, led by the Council for State Governments and funded by the Bureau of Justice Assistance, which also aims to improve accurate, accessible, and actionable data across the justice system. An example of a statewide data collection improvement effort is Washington's program focusing on use-of-force in law enforcement, initiated in 2021. See Engrossed Second Substitute Senate Bill 5259, http://lawfilesext.leg.wa.gov/biennium/2021-22/Pdf/Bills/Session%20Laws/Senate/5259-S2.SL.pdf#page=1.

the logistical barriers mentioned above, the lack of consistent standards and the confusion about race and ethnicity categories pose a critical challenge. This can lead to incomparable datasets across agencies that collect data and to discrepancies between how people identify and how they self-describe within the confines of provided categories.[7] Keeping these limitations in mind, imputing race and ethnicity data from a source such as the U.S. Census is a promising strategy when good quality administrative data on self-reported race and ethnicity is not available.

Strategies for Imputing Race and Ethnicity

Imputation is a method that involves substituting missing data with data from a similar source that can be used as a proxy. It is applied in order to "assign" race/ethnicity information to a dataset by estimating probabilities from other datasets. With either the names, the address-based information, or both, individuals can be assigned a certain race/ethnicity or, better, the probability of possessing that trait. Estimates derived from these data sources can help researchers define race and ethnicity for use in disparity analyses.[8] This paper explores the legitimacy of the approach for use by courts.

Using other recorded personal information such as surnames, common first names, and geocoded addresses from administrative data files, one can impute aggregate race and ethnicity information for a dataset.[9] The assumption behind geocoding is that neighborhoods tend to be racially homogenous for certain races and ethnicities, especially in more segregated regions in the United States. Racially targeted policies related to redlining, public housing, taxation, and labor unions after Reconstruction and through much of the Twentieth Century led to segregated housing patterns throughout the United States.[10] The geocoding technique exploits persistent segregation to impute race. That is, the probability that someone is of a certain race can be constructed by establishing the racial composition of the geographic area in which the individual resides. This is done by associating the individual's address with specific coordinates, a census tract, or simply a ZIP code. Studies using this method alone have often found that it is most accurate when identifying Black and White individuals, but rather inaccurate for identifying Native American and multiracial individuals. It may also lack accuracy in regions with lower residential segregation.[11]

Similarly, certain names are more likely to belong to an individual of a certain race/ethnicity than another. Reference surname lists with the most common names for Hispanics/Latinx and Asians are available through the U.S. Census and other sources, and other reference name lists continue to be created. This

[7] Kathryn Genthon & Diane Robinson, *supra* note 2; National Center for State Courts, *Racial Justice Organizational Assessment Toolkit for Courts: Part II.2. Collecting Administrative Data on Race and Ethnicity*, (Forthcoming).

[8] Jenna R. Sablan, *Can You Really Measure That? Combining Critical Race Theory and Quantitative Methods*, 56 AM. Educ. Res. J. 178 (2019).

[9] Allen Fremont et al., *When Race/Ethnicity Data Are Lacking*, 6 Rand Health Q. 16 (2016).

[10] *See, e.g.*, Richard Rothstein, The Color of Law: A Forgotten History of How Our Government Segregated America (2017).

[11] Marc N. Elliott et al., *A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity*, 43 Health Serv. Res. 1722 (2008); Fiscella and Fremont, *supra* note 2.

method has been shown to estimate Hispanic/Latinx proportions relatively effectively.[12] It is, however, difficult to obtain lists with a utility for all races including White and Black. Other limitations are the temporal or regional differences between reference lists and the estimation sample. Often, names can also be misleading proxies for race, given that surnames are frequently adopted through marriage or reflect only the patrilineal descent.[13]

To overcome the limitations each of the methods alone faces, researchers developed hybrid imputation methods, such as the RAND Corporation's Bayesian Improved Surname Geocoding (BISG) method. This method first uses Census surname lists for six mutually exclusive categories — Hispanic, White, Black, Asian and Pacific Islander, American Indian/Alaska Native, and Multiracial to calculate the probabilities of a person's race given a surname. This probability is then updated by multiplication with the probability of census block residence given a certain race.[14] Several validation studies working in the realm of health care, mortgage lending, and voting noted the relative accuracy and strong correlation between these inferred probabilities and self-reported race/ethnicity.[15] Although the extent to which this method is successful appears to depend somewhat on the sample and geography in question, the evidence indicates that this is a generally promising approach and can likely be applied in other fields such as the justice system.[16]

Nevertheless, it is typically discouraged to use indirect imputation techniques to assign race and ethnicity (or any demographic characteristic) to an individual. What is known as the ecological inference fallacy consists of thinking that relationships or characteristics observed for groups also hold for individuals in those groups.[17] Thus, while we can estimate the *probability* of someone from a certain neighborhood being of a certain race, we cannot predict their race. Instead, the technique can be used to estimate the aggregate demographic composition of a sample. In study designs where the individual is the unit of analysis, the racial characteristic expressed as a probability can be used for outcome or disparity analyses.

[12] Bernard Grofman & Jennifer Garcia, *Using Spanish Surname Ratios to Estimate Proportion Hispanic in California Cities via Bayes Theorem**, 96 SOC. SCI. Q. 1511 (2015).

[13] Pablo Mateos, *A review of name-based ethnicity classification methods and their potential in population studies*, 13 Popul. Space Place 243 (2007).

[14] Elliott et al., *supra* note 3.

[15] Dzifa Adjaye-Gbewonyo et al., *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*, 49 Health Serv. Res. 268 (2014); Fremont et al., *supra* note 9; Elliott et al., *supra* note 11.

[16] For instance, two validation studies found that they could predict race successfully based on the correlation to the self-reported dataset. However, the strength of the correlation varied, ranging from 0.7 to 0.96, see Elliott et al., *supra* note 3; Fremont et al., *supra* note 9; Arthur Baines & Marsha Courchane, *Fair Lending: Implications for the Indirect Auto Finance Market*, (2014), https://www.crai.com/insights-events/publications/fair-lending-implications-indirect-auto-finance-market/ (last visited Jul 16, 2022). Newer iterations of the hybrid method also incorporate first names, which increases the ability to identify Black individuals, see Ioan Voicu, *Using First Name Information to Improve Race and Ethnicity Classification*, 5 Stat. Public Policy 1 (2018).

[17] Philip Sedgwick, *Understanding the ecological fallacy*, 351 BMJ (2015), https://www.bmj.com/content/351/bmj.h4773.

Further Considerations for Employing Imputation Methods to Infer Race and Ethnicity

Race and Ethnicity Categories

- *Which race/ethnicity categories are used for imputation should depend on the geographic and demographic context and the purpose of the study. Researchers must habitually balance agency preferences based on the standard categories of a given jurisdiction with the limitations posed by the available reference data to be used for imputation. Recognizing the implications of categorization differences will lead to a more nuanced and trustworthy interpretation of the results.*

Racial/ethnic identities and the categories to measure them constantly evolve and are dependent on the cultural context. The methodology of the U.S. Census Bureau, a typical source of reference data, has also changed over the years.[18] One significant methodological change (and ongoing debate) involves whether to collect race and ethnicity information in a single data field or to collect each separately. Another pertinent question is whether to use mutually exclusive or multiple selected categories. The latter would account for the multidimensionality of an individual's racial/ethnic identity and recognize the ever-changing nature of these concepts. Using complex categorizations, however, produces challenges for comparisons and quantitative analyses and may create groups too small for statistically significant estimates.[19]

Imputing Data That Will Be Comparable

- *Before imputing race/ethnicity data, researchers must consider the compatibility of the imputed data with the data that will be used for comparison in any analysis of the imputed data.*

The frequent goal of imputation is to compare these data with other population datasets. As there is currently no one universal standard,[20] different administrative systems may employ different classification systems or use similar terminology but define the categories differently.[21]

This poses a challenge when attempting to compare the imputed data with those of other agencies, states, or U.S. Census Bureau information. For instance, if Census data is used for imputation, it should be expected that the "Other" response category will likely be inflated, as Hispanics/Latinx whose only "race" identity is

[18] Samantha Viano & Dominique J. Baker, *How Administrative Data Collection and Analysis Can Better Reflect Racial and Ethnic Identities*, 44 Rev. Res. Educ. 301, 311 (2020).

[19] To overcome this challenge, it might be advisable to focus only on those racial/ethnic groups that are relevant to the inquiry, as determined by the specific decision-making context and purpose of the study. Not all identities chosen in a multi-select measure will be equally salient to the individual. See, Kyle L. Marquardt & Yoshiko M. Herrera, *Ethnicity as a Variable: An Assessment of Measures and Data Sets of Ethnicity and Related Identities**, 96 Soc. Sci. Q. 689 (2015). Another option is to map specific categories into broader categories, see Kathryn Genthon & Diane Robinson, *Collecting Race & Ethnicity Data*, (2022), Court Statistics Project Brief, https://www.courtstatistics.org/__data/assets/pdf_file/0036/69678/Race_Ethnicity_Data_Collection_4.pdf (last visited Sep 5, 2022).

[20] NODS recommends the two-question approach to race and Hispanic/Latinx ethnicity, comparable with the U.S. Census categories. Also see Court Statistics Project Brief, Kathryn Genthon & Diane Robinson, *supra* note 2.

[21] Viano and Baker, *supra* note 18 at 308.

"Hispanic" or "Latinx" may only select "Hispanic" within the ethnicity question and will likely choose "other" as race. Such imputed data can then only be imperfectly compared to administrative datasets that use a distinct "Hispanic" or "Latinx" category in a multi-select measure.[22]

Choosing Appropriate Geographic Units for Geocoding

- *Researchers must decide between more precise but more complicated geocoding based on small geographic units, and the more resource-efficient and simple but less precise geocoding based on ZCTAs.*

The success of using geocoding to impute race/ethnicity depends on relatively segregated residential areas. As census blocks, block groups, or tracts constitute demographically more homogenous units than ZIP Code Tabulation Areas (ZCTA), the accuracy of the estimates tends to be greater when geocoding is performed at smaller geographic levels.[23] Each record needs to be precisely geolocated within Census tracts or blocks using mapping software that may not be readily available.

One of the simplest imputation techniques is to infer the probability of an individual's race based on their ZIP code alone. This is not dependent on mapping software and does not involve additional steps using name reference lists. Nevertheless, there are caveats. ZIP Codes developed by the U.S. Postal Service are not geographically defined areas, but instead reflect postal distribution routes. The U.S. Census Bureau developed ZIP Code Tabulation Areas (ZCTAs) that approximate the geographically defined boundaries for each ZIP Code. However, they are not strictly analogous, and there will not be a ZCTA match for each ZIP Code.[24] If many records without a match have to be excluded from the analysis, the estimate of the racial/ethnic distribution in the dataset will be skewed if a certain group is more prevalent in the excluded records.

Furthermore, ZCTAs do not nest within other geographical units used by the Census and do not correlate to legal or municipal areas. In fact, ZCTA's frequently cross county boundaries and in some instances even state borders. Finally, ZCTA's greatly vary in size. This has important implications for the imputation and comparison of race/ethnicity data. First, race/ethnicity probabilities that are inferred based on ZCTA populations and then aggregated to the county level to compare with the proportions of county populations risk imprecise aggregation and a comparison that is not truly based on the same geographical area. Second, as the inferred probabilities are based on ZCTA populations of various sizes, and the aggregation then involves averaging these percentages across each county, each average will be based on either more or less precise probabilities depending on the geographic unit sizes involved.

[22] *Id.* at 313. For more details on collecting race/ethnicity data that is compatible with a comprehensive set of standard categories and on discrepancies that arise when the compatibility is imperfect, see the Racial Justice Organization Assessment Toolkit for Courts' *Supplement to Part I.1 & I.2: Administrative Data on Race and Ethnicity*, National Center for State Courts (forthcoming).

[23] Nancy Krieger et al., *Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project*, 156 AM. J. EPIDEMIOL. 471 (2002).

[24] For instance, P.O. Boxes and single-address ZIP codes for large-volume customers such as commercial addresses, government entities, or universities will not have a ZCTA match.

Exploring the Viability of the Strategy for Disparity Analyses

There is little research conducted to date on how the imputation methods perform when using inferred race/ethnicity in analyses. These techniques always produce some error, and this error might be systematic.[25] Even if the inferred records overall are comparable to the true population, analysis outcomes, such as disparities, may be over- or underestimated. And if there is significant and systematic error in the imputed demographics, any subsequent analysis based on imputed data will likely produce different results than the same analyses based on self-reported data. Researchers must understand the limitations of their inferred data to know the validity of their disparity analyses. Moreover, they will need to provide court leaders and policymakers with the appropriate context if used to inform decision-making. The case study presented here explores the validity of the geocoding imputation method when examining the representativeness of jury pools.

Case Study: Representativeness of the Master Jury List in Tennessee

Impartial juries, as guaranteed by the U.S. Constitution, require that the jury pool reflect the demographic composition of the geographic community within the court's jurisdiction. The master jury list is the first step in the lengthy process of creating the jury pool from which juries are selected. As such, it is important that the master jury list be as demographically representative as possible to ensure that the resulting jury pool reflects a fair cross section of the community. Researchers assess representation by comparing the demographic composition of the master jury list with that of the jury-eligible population. Ideally, they utilize self-reported race and ethnicity data for such analyses. However, the agencies that maintain juror source lists often do not collect those data.

Inferring the probability of an individual's race based on their ZIP code alone is one of the fastest, most resource-efficient, and most commonly employed imputation methods. In this case study, this method is cross validated with self-reported data from Tennessee's juror source list. Tennessee's juror source list is the list of licensed drivers and state ID cardholders maintained by the state's Department of Safety and Homeland Security. It was shared with the National Center for State Courts (NCSC) in August of 2021 to provide research technical assistance, which included an assessment of the representativeness of the state master jury list.[26] With over 6 million records, this list includes self-reported race/ethnicity for each record. Thus, the original representativeness analyses were based on real, not imputed, data. In this white paper, the results of those original analyses will be compared with the hypothetical results using an imputation method to infer race and ethnicity.

[25] See Arthur Baines and Marsha Courchane, *supra* note 16, who found non-random error in the estimates related to factors such as income levels in the population.

[26] Paula Hannaford-Agor, Miriam Hamilton & Erika Bailey, *Eliminating Shadows and Ghosts: Findings from a Study of Inclusiveness, Representativeness, and Record Accuracy in Master Jury Lists and Juror Source Lists in Three States*, (September 2022), https://www.ncsc-jurystudies.org/__data/assets/pdf_file/0025/82681/Master-Jury-List.pdf.

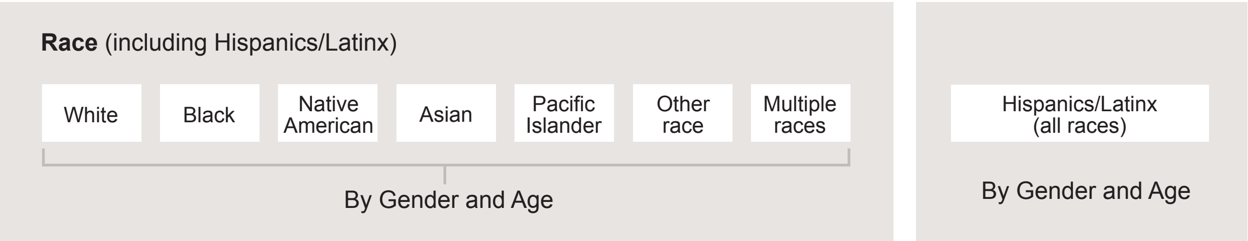
To facilitate comparison, first race and ethnicity for each record were inferred using population data from the U.S. Census Bureau for each ZCTA.[27] The inferred race/ethnicity probabilities were then aggregated to the county level. Unfortunately, the racial classification used for the Tennessee list does not adhere to the U.S. Census classification system. The U.S. Census defines Hispanic/Latino as a separate ethnicity category, and Hispanics may be of any race. On the other hand, Tennessee's list includes Hispanic as a race category exclusive to all other races. Additionally, in the Tennessee list, "other" describes all races other than White, Black, Asian, Hispanic, and Native American, whereas the U.S. Census distinguishes Native Hawaiian/Pacific Islander and "multiple races" as additional categories to "other." The disparity analyses reported in this white paper focus on the following groups: White, Black, and Hispanic/Latinx.

Testing Three Ways to Use U.S. Census Race and Ethnicity Data for Geocoding

Disparity analyses and other demographic comparisons are often limited even when the race data is known if the racial classification systems used between data sources differ from one another.[28] To identify the best reference dataset among the available options, three datasets are tested. Each set is obtained from the U.S. Census and presents a slightly different tabulation of the same race and ethnicity data. The three classification methods with the included race/ethnicity categories are described below.

- **Version 1: Hispanic-inclusive race percentages of the adult population.**[29]
  This version uses the probability that an individual is Hispanic/Latinx as an additional ethnicity category.

Census Tabulation Version 1/3:



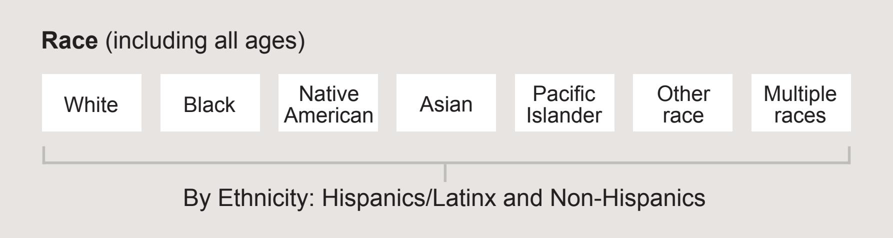
[27] The source for all data utilized is the American Community Survey (ACS), 5-Year Average, for the Years 2015-2019. It should be noted that, unlike the decennial census, the ACS estimates population totals. Particularly for very small populations, such as minorities in small ZCTAs, the margin of error can be relatively large. ZCTAs reflect the geographic boundaries of residential ZIP Codes.

[28] This also leads to a genuine difference at the level of individual data collection. For instance, an individual may identify both as racially White and ethnically Hispanic/Latinx if given the option, but faced with the choice between the two, may identify as either. Those who would prefer to select "Hispanic" as race, may opt to select "other."

[29] Version 1: U.S. Census Table B01001, Version 2: U.S. Census Table B03002; Version 3: U.S. Census Table B01001.

- **Version 2: Hispanic-exclusive race percentages of the total population.**
  This version is based on another tabulation of the same data that breaks down each race category into Hispanics/Latinx and non-Hispanics.

Census Tabulation Version 2:

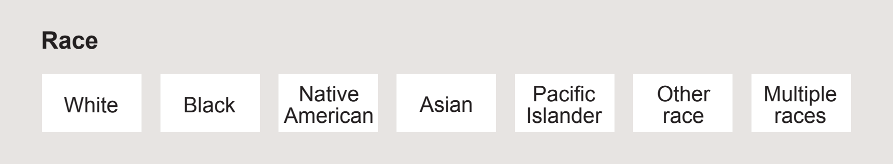


Race (including all ages)

White | Black | Native American | Asian | Pacific Islander | Other race | Multiple races

By Ethnicity: Hispanics/Latinx and Non-Hispanics

Tennessee List Tabulation:



Race

White | Black | Native American | Asian | Pacific Islander | Other race | Multiple races

This version may allow for a cleaner comparison with the Tennessee classification system. However, this version excludes all Hispanics/Latinx from the race categories, including those who might have identified more strongly with another race if asked to choose. The other disadvantage is that these inferred demographics are based on the total population, including children, whereas the Tennessee master jury list contains records of adults only.[30]

- **Version 3: Hispanic-inclusive race percentages of the total population.**
  Unlike Version 1, which also uses Hispanic-inclusive percentages but infers demographics based on the adult population, this version infers demographics, like Version 2, based on the entire population, including children. It is hypothesized that this may produce greater error than Version 1's inclusion of Hispanics/Latinx in the race categories.[31]

[30] The original list contained individuals under 18. Because Tennessee law requires that citizens be age 18 or older to qualify for jury service, the list was cleaned and minors were excluded based on the date of birth recorded in the list.

[31] For Hispanics, Versions 2 and 3 produce identical results, as both are based on total populations, and the exclusion or inclusion of Hispanics from the other race categories only affects those categories, not the Hispanic percentage itself.

Methodology

Evaluation of inferred demographics

To evaluate the validity of the imputation method, the three versions of inferred demographics are evaluated and compared in terms of accuracy in a sequence of steps. First, the question of how closely the estimates each version produces match those derived from the true, self-reported race/ethnicity is explored via the Pearson's correlation coefficient.

The next step is to explore the differences between inferred and true county averages of race/ethnicity and whether they are statistically significant with a Wilcoxon Signed Rank Test. A comparison of the statewide averages sheds additional light on accuracy and whether the versions over- or underestimate the prevalence of each race/ethnic category in the dataset. Ideally, inferred race and ethnicity information would be perfectly correlated with self-reported race and ethnicity. However, as noted earlier, inference is inherently imperfect, so there will be some difference between the data sources. If the difference between datasets is statistically significant, using an inferred data source for disparity analysis will likewise lead to results that differ from results based on self-reported data.

A related question is whether these differences might be affected by other factors, such as the overall size of the racial/ethnic group in a county. This potential relationship is explored via visual analysis of scatterplots and additional significance testing for the subgroups of each sample.

While it is important to evaluate the performance of the inference method for each race/ethnicity proportion individually, the accuracy of the overall distribution also needs to be considered. The question is how the distributions of race/ethnicity in the inferred versions compare with the true distribution and which version produces the most accurate overall distribution. Following previous research, this is answered by summarizing the distributional differences across race/ethnic groups and counties by computing the weighted average of the deviations, with weights given by the true proportion in each category.[32]

Evaluation of derived disparity measures

While the first part of the analysis focuses on the accuracy of the estimates and overall distribution of inferred race/ethnicity, a pressing remaining question is whether the differences are large enough to substantially skew subsequent analysis using the inferred demographics as data source. The second part of the evaluation thus focuses on the bias introduced by the imputation method. To assess how a representativeness analysis would be affected by the estimation errors, disparities are calculated between each set of inferred race/ethnicity percentages and the demographic true composition of the relevant

[32] The evaluation strategies used here largely follow previous research on imputation method performance, see Elliott et al., *supra* 11; Elliott et al., *supra* note 3; Voicu, *supra* note 16; Arthur Baines and Marsha Courchane, *supra* note 16. However, there are two main adjustments. First, the unit of analysis here is the county aggregate of inferred individual probabilities rather than the individual record. Hence, correlational analyses, tests of significance, and the averages of the distributional differences are here all computed based on the county percentages. Second, this study goes a step further by not only evaluating the accuracy of the inferred demographics but also the bias that might be introduced through imputation through a comparison of disparity analysis results (see similarly, Voicu, *supra* note 16.)

community, in this case, the jury-eligible county population.[33] These disparities can then be compared with disparities derived similarly from self-reported data. The demographic composition of the jury-eligible population is once again based on data collected by the U.S. Census Bureau.[34]

For the comparison, each of the steps for evaluating the inferred data is repeated in a similar fashion with the disparity measures to determine how greatly the disparity results differ. The final question to ask is, “How much estimation error is too much, and how great of a skew in the outcome results is still acceptable to validate the imputation method?” Unfortunately, this last question depends on the thresholds relevant for the disparity analysis to be performed, which will differ based on the study question and context.

Courts often measure jury pool representativeness via absolute and comparative disparity, and common thresholds are 10% for absolute and 50% for comparative disparity (see Box below). These thresholds are used here to assess in how many counties the results of each version produce disparities above the typically acceptable threshold. Whether or not the same counties are flagged for exceeding the thresholds (raising representativeness concerns) helps to determine how valid the imputation methods might be.

Measuring Disparity of Jury Pools: Absolute and Comparative Disparity

Absolute disparity is the numerical difference between the percentage of a racial group in the jury pool and in the community. Comparative disparity, or relative disparity, measures the percentage by which the group members’ number in the jury pool falls short of or exceeds their number in the community.

A criminal defendant can establish a prima facie violation of the fair cross-section requirement when (1) the allegedly excluded group is a distinctive group in the community, (2) the group’s representation in the jury pool is not fair and reasonable in relation to its representation in the community, and (3) the under-representation of the group results from systematic exclusion.[35] While there are no specified thresholds for absolute or comparative disparity, the consensus emerging in case law is that absolute disparities above 10% and comparative disparities above 50% may be sufficient to show a prima facie violation of the fair cross section requirement.

[33] For purposes of a feasible disparity analysis, a “jury eligible” person is defined here as an adult citizen living in the county served by the court, although there are additional criteria that determine jury eligibility.

[34] Tables B05003, B05003A-I, ACS 5-Year Estimates Detailed Tables, 2019. These race-specific tables contain estimated totals by age and citizenship status for each county.

[35] As described by the U.S. Supreme Court in *Duren v. Missouri*, 439 U.S. 357 (1979).

Findings

Comparing Demographics in Versions 1,2, and 3

Correlations

All three sets of inferred demographics were strongly correlated with actual self-reported race/ethnicity data in the Tennessee juror source list (p<.01). Details are listed in Appendix 1. While differences between versions were minor, the most accurate estimates, as judging by slightly greater correlation strength, were achieved for Hispanics/Latinx, Whites, and Blacks by using inferred demographics based on total populations (Versions 2 and 3). This is somewhat counterintuitive given the all-adult master jury list. For Whites, the correlation to self-reported demographics was strongest when additionally excluding Hispanics from the race categories (Version 2). For Blacks, this did not make a difference.

Differences between inferred estimates and self-reported data

Comparing the statewide averages of the race/ethnicity probabilities, as shown for Whites, Blacks, and Hispanics/Latinx in Table 1, presents a slightly different pattern. Each version leads to averages that are closer to the self-reported average for one race, but not another, and there is no clear pattern of over- or under-estimation. Across versions, the inferred county percentages had a notably smaller range for Whites and Hispanics than the true demographics had.[36]

Table 1: Comparison Between Infererred and Self-reported Race & Ethnicity, averaged across 95 TN counties

	Median for White	Median for Black	Median for Hispanic/Latinx
Self-Reported	**92.5%**	**2.8%**	**1.8%**
Version 1	93.5%	3.4%	2.2%
Version 2	89.5%	3.4%	2.8%
Version 3	92.6%	3.4%	2.8%

*Differences in race/ethnicity prevalence between the inferred and self-reported samples for this category are significant at the 1% level.

Although all three versions produced race and ethnicity estimates that are very close to the self-reported data, as indicated by the strong correlation between them, statistically significant differences between them can indicate important variance that warrants further analysis before determining suitability for use. Testing for statistical significance with a Wilcoxon Signed Rank Test reveals that for some races and inference versions the difference to the self-reported data is indeed statistically insignificant.[37] Detailed test results are provided in Appendix 1, Table B.

[36] A full table with summary statistics including interquartile ranges is shown in Appendix 1, Table A.

[37] It may be noted that due to the large sample, it might be expected that even minor differences would be statistically significant, as was the case in previous research for all race categories and methods investigated (Elliott et al., *supra* note 11; Voicu, *supra* note 16). Here, the finding that there were instances where the differences lacked significance is thus quite interesting.

In summary, for Whites, it is Version 3 that produced a statewide median closest to the true median and inferred percentages that are insignificantly different from the true data. The imputed percentages for Blacks are indifferent from the true list percentages, regardless of whether Hispanic-inclusive race categories and/or adult-only populations were used, although each version of imputed Black percentages was on average overestimated. For Hispanics/Latinx, all inferred percentages were also on average higher than the true data. Yet, Version 1 produced percentages that are most accurate, and this is also the only version for which the differences to the self-reported percentages are insignificant.

Based on these findings, using adult populations as reference data (Version 1) seems best suited for disparity analysis where Hispanics/Latinx are concerned. For Whites, on the other hand, total, Hispanic-inclusive populations (Version 3) led to the best estimates. All versions were successful in estimating Black proportions.

County-specific differences based on racial and ethnic group proportions

Comparing the size of the differences between estimates and true race/ethnicity percentages by county reveals that there are systematic county-to-county variations. The proportional size of the racial group in the total county population affected estimation errors for Whites, Asians, and Hispanics/Latinx. For Whites, where including Latinx in the count led, on average, to overestimated imputed race percentages, this is particularly true in counties with White populations below 90%, which applies to about 40% of all 95 counties. Appendix 2 Figure A shows the differences between Version 1-inferred and self-reported demographics by county for Whites. Results in those counties with above 90% White populations are much more accurate, i.e., the differences are below five percentage points and statistically insignificant.[38] Thus, for counties with proportionately large White populations, including Hispanics/Latinx in the count and using adult-only populations led to a better estimate.

Furthermore, counties with proportionately larger Asian populations produced greater underestimations of the proportion of Asians in the county's inferred race dataset (see Appendix 2 Figure E). For counties with overall larger Hispanic/Latinx populations (above 3% on the master jury list, 29 counties), the true Hispanic percentages were significantly higher (median=5.0%) than the inferred percentages (median=4.5% in Version 1). For counties with less than 3% Hispanics/Latinx on the master jury list (66 counties), the true list percentages were significantly lower (median=1.4%) than the inferred percentages (median=1.9% in Version 1). Appendix 2, Figure F shows this pattern in more detail. Although the overall difference was found to be insignificant, the overestimations for counties with smaller Hispanic/Latinx populations are balanced in this study by the underestimations in counties with larger Hispanic/Latinx populations.

Caution may thus be warranted when using the inference results for individual counties. As the findings indicate, counties with proportionately large White populations produced better estimates for Whites. For Asians, larger Asian populations were underestimated to a greater degree than small county populations. For Latinx, the very small populations were underestimated even more, while relatively larger populations were overestimated.

[38] See details in Appendix 2.

Accuracy of the overall distribution of estimates

To evaluate the relative accuracy of the overall racial/ethnic distribution in each version, the distributional differences between the sets of inferred and those of the self-reported proportions were summarized across categories. The deviations were each weighted by a county's true proportions of the racial/ethnic group, and the statewide estimation error averages are shown in Appendix 3.

Comparing these, Version 3 is shown to be the most accurate overall with the smallest statewide overall deviation from the true distribution. Based on the findings described above, this version indeed performed especially well with Whites and Blacks, only insignificantly overestimating their proportions. On the other hand, it significantly overestimated the prevalence of Hispanics/Latinx. While all versions performed well for Blacks, Hispanics are the only group for which another dataset, Version 1, would be better suited for disparity analysis given the findings, as the Hispanic proportion is only insignificantly underestimated in Version 1.

How big the estimation error can be before it is considered too big, may, however, depend on the purpose of the analysis using inferred race and ethnicity. The findings presented thus far help researchers realize the probability for overestimation or underestimation of each race category depending on the data source for imputation. This is important when utilizing the inferred estimates to derive disparities that are then measured against certain thresholds for concern, given the potential policy or legal consequences of a representativeness evaluation. Whether the imputation-derived disparities are more or less likely to exceed such thresholds compared to disparities derived from known data is addressed in the final section below.

Disparity Findings Using data from Version 1,2, and 3

Focusing on the question of how great the deviation can be before dismissing the ZIP-code-based inference method as a viable option, the disparity measures calculated based on inference are evaluated and compared here with those calculated based on self-reported race and ethnicity.

Differences between disparities derived from estimates and true disparities

It is notable that in most cases the master jury list appears more representative with inferred race demographics and the distribution of disparity measures across counties is more homogenous than the results based on the true records would have indicated. In most instances, the disparities have a mean closer to zero, a smaller range, and fewer outliers compared to the measures derived from self-reported race percentages. Nevertheless, there are also instances where the disparity is overestimated, such as for Whites in Version 2. A comparison of summary statistics is shown in Appendix 4.

The relative accuracy of the estimated disparity measures mirrors the results of the comparison of demographic proportions between versions.

Version 1. Using estimates from Hispanic-inclusive adult populations led to derived disparities that:

- are significantly smaller for Whites, making the master jury list appear more representative;
- show a very close estimation of the Hispanic/Latinx overrepresentation. (The slight overestimation is statistically insignificant); and
- show a very close estimation of the Black underrepresentation. (However, the true underrepresentation on the list is accurately expressed by the disparity measures regardless of the version, and any deviation is statistically insignificant.)

Version 2. Excluding Hispanics/Latinx from the inferred race demographics led to:

- statistically significant over- or underestimation for disparities of all racial groups except Black.

Version 3. Including Hispanics/Latinx in the racial count but deriving estimates from total populations:

- works well for Whites, where the slight underestimation of White underrepresentation on the list is statistically insignificant; but
- does not perform as well for Latinx, who would have seemed significantly more overrepresented on the list than they are.

Accuracy of the overall distribution of disparities across counties

Assessing the overall severity of the disparity inaccuracies across all races in each version, it is again Version 3 that provides derived disparities that show the smallest statewide overall deviation from the true disparity distribution, as shown in Table F, Appendix 4. Yet, comparing the summary statistics of disparities for each race and testing for significance of the deviations from the list-based, true disparities, Version 3 only truly performed well for Whites and Blacks. While Black disparities measured very close to the true disparities in all versions, unfortunately, there was no version that worked well for inferring both Whites and Hispanics/Latinx.

Whites are generally underrepresented on the master jury list of this case study. Using adult-only populations for inferred race and ethnicity data led to Whites not appearing as underrepresented as they should. To reflect the true underrepresentation more accurately, using the total populations (Version 3) for inference helped in this case. As the White birth rate is lower than that of most other races,[39] the comparatively small proportion of White children in the total-population estimate relative to that of other races will lead to a smaller estimated proportion for total Whites than when using adult-only populations for all races. However, Hispanics/Latinx, who have higher birth rates, are always overrepresented here.[40] Using total population-based inferred data only exaggerates this overrepresentation.

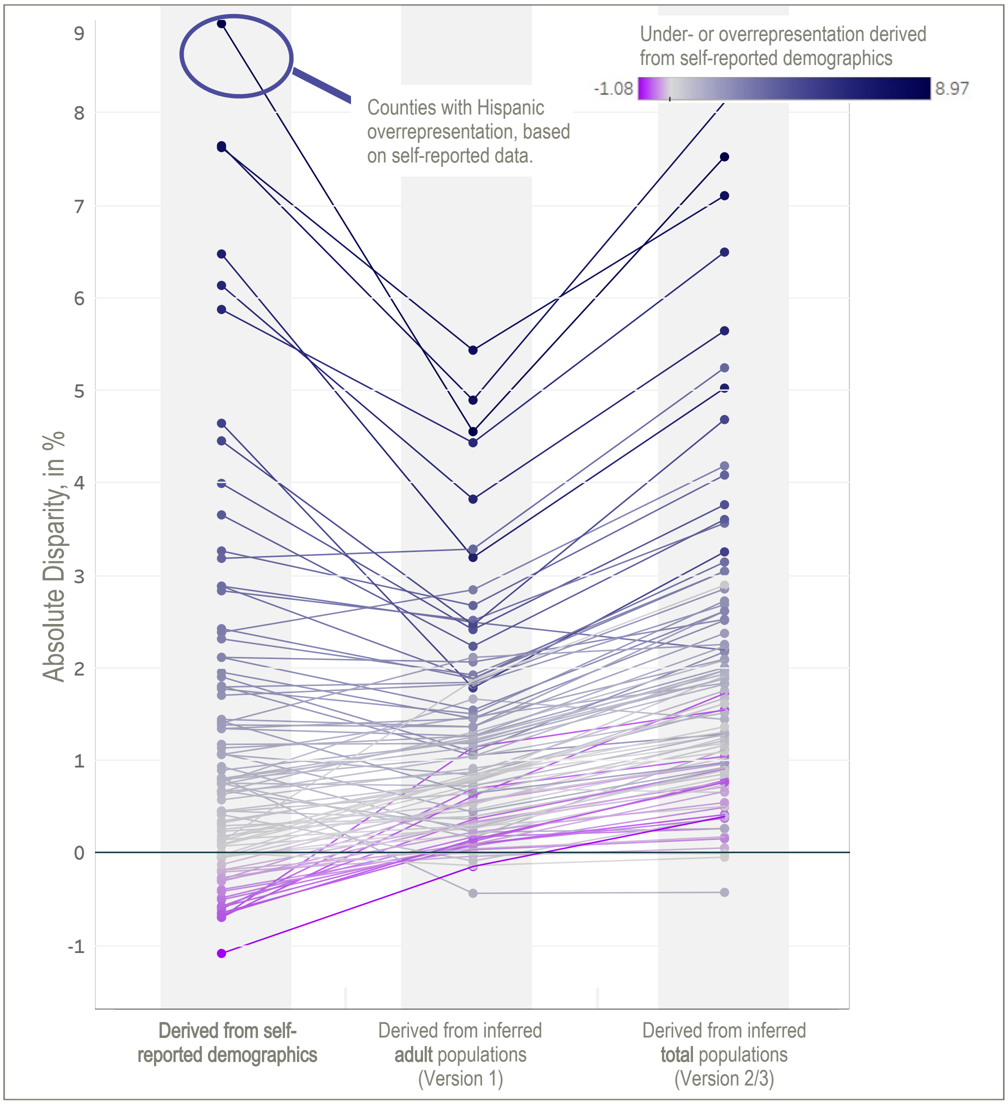
[39] Joyce A Martin, Brady E. Hamilton & Michelle J.K. Osterman, *Births in the United States, 2020*, Data Brief No.418, 1–8 (National Center for Health Statistics 2021), https://www.cdc.gov/nchs/data/databriefs/db418.pdf.

[40] Generally, the original overrepresentation of Hispanics and Asians on the master jury list based on self-reported race and ethnicity can, in many instances, likely be attributed to high non-citizen rates for these groups. The source list from which potential jurors are summoned in Tennessee includes non-citizens as citizenship status is unknown at this early point in the jury management process.

Results for disparity versions by county

A court with county-wide jurisdiction will summon jurors from the jury-eligible population residing within county borders, and thus each county's list would need to be assessed individually in terms of representativeness. The parallel charts in Figure 1 show Hispanic disparities by county for those counties with Hispanic eligible populations above 2%.

Figure 1. Disparities for Hispanic/Latinx Between Juror Source List and County Population in 95 Counties, by Data Source



As shown, both versions based on estimates generally led to higher disparity values – in this case, an overestimation of the overrepresentation of Hispanics/Latinx. However, this is more extreme in the version based on total populations. Nevertheless, this is the version where for those counties with larger absolute disparities (dark blue), the disparities derived from total population estimates are most accurate. While aside from these higher disparity outliers the middle version seems to produce generally closer estimates, the small range of these estimates is clearly notable. The true disparities have a much greater range and even include counties with Hispanic/Latinx underrepresentation (i.e., absolute disparity values below 0%).

This same result is found for other races too, such as is shown in Appendix 5, Figures G and H for Whites and Blacks, where the inference-derived disparities have a much smaller range across counties than the true disparities do. This is especially evident for the versions including Latinx in the race count.

Looking at individual counties, the true extent of White underrepresentation in the most extreme cases is only picked up when using total populations and Hispanic-exclusive estimates, despite the statewide overall performance being better with the Hispanic-inclusive version 3. Furthermore, none of the versions can reproduce the original absolute disparity findings for the few counties with White overrepresentation above 3% absolute disparity. For Blacks, this analysis has so far shown that proportions and derived disparities can be relatively accurately predicted. However, for the few outlier counties where the true disparities are at potentially concerning levels of underrepresentation, the same counties would not show similarly high disparities when imputing race.

Quantifying the imputation effect on disparity results

The most relevant effect of the imputation can be found when assessing in a representativeness analysis how often and in which counties the disparities exceed certain specified thresholds — those commonly-held thresholds that might raise concerns and potentially warrant action to be taken. Ideally, any analysis that is based on inferred demographics would identify potential problems for the same counties as an analysis based on self-reported demographics would have identified. Likewise, it would not identify problems that would not have been flagged if the true racial and ethnic proportions on the list were known.

Analyzing the master jury list, a representativeness analysis aims to identify counties in which over- or underrepresentation of a distinctive group exceeds 10% absolute disparity and 50% comparative disparity. Here, all analyses based on inferred race and ethnicity led to an overall more positive picture, with fewer counties identified in the more extreme disparities ranges than should have been identified. This is true for all racial groups. Appendix 6 details the results and shows the distribution of disparities within ranges of over- and underrepresentation depending on the list version used for calculating the disparities for Whites, Blacks, and Hispanics/Latinx.

For Blacks, the original analysis had identified six counties with clear underrepresentation on the jury list — with above 50% comparative disparities (CD).[41] Only one out of these six would have been identified when using estimated demographics. Considering Black overrepresentation above 50% CD, the estimated versions did not identify the one county correctly where such overrepresented was noted based on self-reported data. They did, however, incorrectly flag one to two new counties each with seemingly great overrepresentation above this threshold where such a disparity did not exist based on self-reported data. Similarly, the lesser underrepresentation ranges (20% - 50% CD) included about 6.5% of the counties for which measures could be calculated.[42] The number of counties identified in these ranges by the three inference versions range between 3% and 8% of total counties. However, again, these were not necessarily the same counties that should have been identified in this range.[43]

[41] Comparative disparity is often a better measure for very small populations.

[42] For Blacks, this was 62 counties with eligible populations of this race/ethnicity >2%.

[43] The adult-only estimate does not include any of the original counties in this range, and the two total-population estimates identify only half of the original counties in this range but include other additional counties.

Whites, never overrepresented based on the original analysis, but severely underrepresented in three counties, are shown with the same threshold-exceeding underrepresentation in the same counties only when excluding Hispanics from the estimates (Version 2). Hispanics/Latinx were strongly overrepresented in most counties according to the original disparity analysis. These counties were each identified as such, or nearly so, in each inference-based analysis. Each version, however, falsely indicated apparent severe Latinx overrepresentation for at least one additional county, in which, according to the original analysis, Latinx were even underrepresented on the list. Thus, for these few counties, using inferred ethnicity led to results that were opposite to the original findings, reversing the direction of the finding from minor underrepresentation to severe overrepresentation.

Assessing total findings, Hispanics/Latinx were best inferred when based on adult populations, not surprisingly, as the original list included adults only. Whites, on the other hand, were on average best inferred when using total population demographics and including White Hispanics in the percentage. The Black estimates could be inferred most successfully; they are insignificantly different from the true race proportion no matter whether adult or total populations are used in the estimate or whether Hispanic-inclusive racial counts are used. Two factors contributing to these results are, first, differing birth rates which influence the relative race proportions in the adult population versus the total population. This is important to keep in mind when inferring race and ethnicity for Whites and Hispanics/Latinx in particular, as total-population counts are not a perfect proxy for inferring adult population probabilities for these races. Second, Latinx are, on average, quite overrepresented on this list while Whites are, on average, underrepresented. This fact means that a successful inference method would result in the same overrepresentation finding for Latinx — one that is not exaggerated by including the relatively larger proportion of Hispanic/Latinx children in the estimate — and the same underrepresentation finding for Whites – which a simple address-based inference method appears to underestimate. In the case of this study, including minors in the demographic counts helped to get closer to the true White disparity, skewing the results toward a lower estimate.[44]

Importantly, for the few counties with severe over- or underrepresentation and thus outliers in the original distribution of disparities among counties in Tennessee, imputation was unsuccessful in estimating the true proportions. While not an issue for Hispanics, this was particularly notable for Black underrepresentation. Additionally, even when imputation-derived disparities identified more severe under- or overrepresentation, it was not necessarily for the same counties that are truly in these ranges as discovered by the original analysis. The results did not clearly indicate which version might lead to the correct identification of those counties with potential representativeness issues. An exception is possibly Version 2, which identified the three counties with high White underrepresentation.

[44] For this analysis, American Community Survey data was used from approximately the same year (2019) as the year when the records of the master jury list were created, and, as birthdates were known, minors could be excluded. However, this might not always be possible. Given the effect that including minors in the reference data can have on the accuracy of the inferred demographics, considering the age composition when choosing reference data is crucial.

Summary of Key Findings

This case study explored whether imputing race/ethnicity via Zip Code-based geocoding is a viable option when using results for disparity analysis, utilizing the records on the Tennessee master jury list as an example. Specifically, this study assessed how close the estimated data were to the true, self-reported data, comparing three imputation source data versions that either excluded or included Hispanics/Latinx in the racial count and either used total or adult population as basis for imputation. Disparities derived from the estimated demographics were then evaluated in terms of their accuracy compared to those derived from the known data.

We conclude that court researchers can successfully impute race and ethnicity demographic information from Census data using ZIP-code-based geocoding, particularly for White, Black, and Latinx groups. However, we also conclude that the accuracy of the imputation method varies across jurisdictions, across race and ethnicity categories, and across the versions of Census data used for imputation. As also shown here, the overall accuracy of the estimated list-wide racial/ethnic distribution does not automatically include accurate proportions for each group individually. Notably, using the imputed data for disparity analysis led to results that suggest a more representative master jury list for Blacks than it truly is, as the outliers with severe under- or overrepresentation were not successfully identified.

The success of the imputation method may be limited to: jurisdictions in which there is a high degree of neighborhood segregation, original high-quality, residence-only address data used for geocoding, a particular reference data source for each racial group (e.g., total Hispanic-inclusive populations for Whites, adult-only populations for Hispanics/Latinx), and a small likelihood of outliers in the dataset to be imputed, i.e., no counties for which the dataset includes a high proportion of individuals of a certain race/ethnicity living in a ZIP code community where that race/ethnicity has smaller proportions. The success of the method may also be influenced by the overall size of a group's population in the community. Estimation errors might be larger when a group's relative size is very small compared to where the group is larger, or vice versa. Similarly, variance in birth rates among racial and ethnic groups may have a different effect on the success of using a particular imputation data source in one jurisdiction compared to another.

Because of these limitations and influences, we strongly recommend that each jurisdiction entertaining the use of imputation methods as a technique for obtaining race and ethnicity data consider the findings from this case study and other research in selecting the parameters of inference methods appropriate to their local jurisdiction and specific population before using the inferred dataset to conduct disparity analyses. Such parameters include the format of the U.S. Census tabulations for various geographical units, Hispanic-inclusive or exclusive race categories, and age groups among others. We suggest further research to establish criteria for the types of jurisdictions and populations for which the methods presented here are most appropriate in order to assess the likely validity and viability of the chosen method for a specific local jurisdiction and population. In the case studied here, Black estimates could be inferred most successfully, except for individual outlier counties, regardless of the census data version used. Yet only a certain version was successful for Whites, and another, less consistently, for Hispanics/Latinx. Thus, while these findings give insight into which census data might be preferable in the case of each race/ethnicity, the findings only apply to the jurisdiction and original dataset of the case study.

Implications for Court Researchers

1. **Not all race and ethnicity categories can be inferred equally well using the same imputation methods.** The evidence presented here confirms prior research that White and Black proportions may be most easily inferable with geocoding methods. This is likely because Blacks, more than other races, often tend to live in racially concentrated neighborhoods.[45] Researchers consistently found relatively poor predictability of multiracial and American Indian/Alaska Native people, and this study confirms this. Here, as in similar research, it might be argued that this caveat to the method might be disregarded as these populations are usually less than 1% of the overall population. However, as demographics are changing and more people identify as multiracial, the utility of the method might decline.[46] Researchers should consider the geographic context of their study and the salient demographic group(s) before deciding on methods and reference data.

2. **Geocoding-based imputation methods will likely fail when the data to be imputed include outliers or where the assumption of neighborhood segregation is not given.** Inferring race and ethnicity based on ZIP codes tends to produce a generally rosier picture, polishing off the rough spots and presenting a dataset as more representative than it is. Here, this was found by assessing results for those counties for which the list's true records do not mirror the community demographics. Using inferred disparities, these counties appeared to have "better" race/ethnicity counts, i.e., proportions closer to those found in the community. This will likely always be the case when the records' race probabilities are being inferred based on the demographics of the ZCTA community. Unless ZCTAs show extremely clear racial segregation, the probabilistic nature of the inferred race and ethnicity proportions will not be sensitive to the unlikely outliers that happen to be on the list. Therefore, when feasible and available resources allow, opting for more precise methods of imputing race might be sensible.

3. **How Latinx are counted can lead to complications, such as imperfect comparability between different classification systems and findings that require complex interpretation.** Here, using a system that includes Hispanics/Latinx in each race category and measures Hispanics/Latinx as ethnicity to infer probabilities that are then validated by demographics that include Hispanics/Latinx only as an exclusive race category led to complex findings and was not ideal. This might be unavoidable, given the limited availability of reference data, as discussed below. However, if possible, carefully considering options given the classification system in use in a local jurisdiction and the available other options that might lead to better comparability and more straightforward results is strongly recommended.

4. **A mixed approach using different imputation reference data sources and methods for different races and ethnic groups might be a promising solution to overcome limitations.**

   a. **Mixed census demographic data sources.** Here, the approach of using total populations and Hispanic-inclusive race categories to infer race yielded the best overall, though not perfect, results. Since this approach only yielded accurate estimates for Whites and Blacks,

[45] Allen M. Fremont et al., *Use of geocoding in managed care settings to identify quality disparities*, 24 Health Aff. Proj. Hope 516 (2005).

[46] Viano and Baker, *supra* note 18.

one interesting possibility might be to settle on using this imputation version for inferring all racial categories but not Latinx. Since Hispanics/Latinx are measured as an additional ethnicity category, it is possible to use a different dataset, such as adults only, specifically for Latinx without affecting the percentages of the other racial groups. In most scenarios, there will not be a validation sample to compare to. Since the comparison data for the jury-eligible population, which is needed for computing the racial disparities, can be taken from the same census tabulation, such a mixed approach might lead to the most accurate race and ethnicity proportions and, subsequently, the most accurate disparity measures.

b. **Smaller geographic units.** Another approach is to infer race based on smaller geographical areas, such as census blocks, block groups, or tracts, which are demographically more homogenous than ZCTA's but require more complicated geocoding processing. For demographically more homogenous units, the accuracy of the estimates tends to be greater as geocoding is performed at more precise geographic levels.[47] This may also mitigate the tendency of the ZIP-code-based analysis to "gloss" over small neighborhood enclaves with a large minority population that happens to be represented on the list but would not greatly influence the ZCTA-wide demographic estimate. Census tracts are equal in population size, which aids in the correct calculation of averages.

c. **Surname analysis for imputing Latinx and Asian proportions.** Particularly if the Hispanic/Latinx or Asian proportion is of concern, using a hybrid method that includes surname lists might be more effective in producing accurate results for these racial groups that do not tend to live in segregated neighborhoods.[48] Such an approach might be particularly advisable when the list of records to be inferred is a smaller sample of the population.[49]

5. **Any imputation method will be limited by the availability of reference data.** In this case, as in other studies, the publicly available tabulation formats that the U.S. Census offers restricted the ways race and ethnicity could be inferred. Ideally, Hispanic-exclusive race categories for adult populations would have been used. This was not possible for this study. However, if the time frame fits, demographic data from the decennial census can be used as it is indeed tabulated in a way that enables researchers to use Hispanic-exclusive race categories for the adult population only.[50]

6. **Even successful imputation will introduce some bias to the disparity analysis.** It is essential to acknowledge that even with a method that has been proven to be valid, the success may vary given the factors mentioned above, and these influences will likely not be mitigated perfectly. Whenever findings of a disparity analysis based on inference are presented, researchers need to be aware of the degree to which the findings are likely more positive than warranted, and that there may

[47] Krieger et al., *supra* note 23.

[48] Elliott et al., *supra* note 3.

[49] The master jury list here presents a fairly inclusive list of the adult population, as an ideal master jury list should be inclusive of every adult citizen within a given geographic jurisdiction. Hence, using a purely address-based approach was more justifiable in this study than it might be in other scenarios.

[50] This is Table P4 of the 2020 Decennial Census. Another advantage to using decennial census data is that it does not rely on weighting to account for small samples of certain racial groups, like the American Community Survey as a sampling-based survey does.

be smaller geographic areas (e.g., counties within states, townships within counties) with severe disparities that are not sufficiently flagged. If an evaluator expects that ground conditions likely look worse than analysis findings suggest due to the inference method, it might be advisable to use a more stringent standard for raising concerns to avoid missing issues. In this case study, this would mean lower than 10% absolute disparity and lower than 50% comparative disparity. At the same time, large relative (comparative) disparity findings should not be misinterpreted as they can easily become excessively large and misleading for populations that are overall very small.[51] Consequently, an insignificant, small difference between the inferred demographics and the true demographic proportions may easily influence the disparity measure enough to even reverse the direction from underrepresentation to overrepresentation and vice versa, even if the demographic proportions are estimated seemingly accurately. Due to the sensitivity of the comparative disparity measure for small populations, any threshold used as a decision standard should be interpreted in a case-by-case manner.

[51] This was the case for Hispanic/Latinx populations in many of the counties presented here.

Technical Appendix

APPENDIX 1

Comparison between inferred and self-reported race and ethnicity

Correlations

For Whites, the correlation is strongest[52] for Version 2 (excluding Hispanics, total populations). For Blacks, Version 2 and Version 3 (using total populations) have just a slightly stronger correlation with self-reported data than Version 1,[53] and this is true for Hispanics/Latinx as well.[54] For Asians, on the other hand, it is Version 1 that leads to a higher correlation than Versions 2 and 3.[55] Using both Hispanic-inclusive categories and adult populations also benefitted the accuracy of inferred Native American[56] counts.

Table A. Comparison Between Inferred and Self-reported Race and Ethnicity, summarized across 95 TN counties

		Mean %	Standard Deviation %	Median %	IQR %
White	Version 1	89.6	10.6	93.5	8.0
	Version 2	86.2	11.6	89.5	8.7
	Version 3	88.7	11.1	92.6	8.1
	Self-Reported	**88.2**	**11.9**	**92.5**	**10.3**
Black	Version 1	7.3	10.2	3.4	7.0
	Version 2	7.2	10.4	3.4	6.5
	Version 3	7.3	10.5	3.4	6.6
	Self-Reported	**7.1**	**10.3**	**2.8**	**6.6**
Native	Version 1	0.3	0.2	0.3	0.3
	Version 2	0.4	0.4	0.4	0.3
	Version 3	0.3	0.3	0.3	0.3
	Self-Reported	**0.2**	**0.1**	**0.2**	**0.1**
Asian	Version 1	0.7	0.8	0.5	0.7
	Version 2	0.4	0.4	0.4	0.3
	Version 3	0.7	0.8	0.5	0.6
	Self-Reported	**1.0**	**1.0**	**0.7**	**0.8**
Hispanic/ Latinx	Version 1	2.9	1.9	2.2	1.7
	Version 2	3.6	2.4	2.8	2.1
	Version 3	3.6	2.4	2.8	2.1
	Self-Described	**2.9**	**2.6**	**1.8**	**2.4**

[52] Pearson's r(93) = .99, p=.000. For Version 1 and Version 3, r(93) = .97, p = .000

[53] Pearson's r(93) = .993, p=.000. For Version 1, r(93)=.992, p=.000.

[54] Pearson's r(93) = .96, p=.000. For Version 1, r(93) = .95, p= .000.

[55] Pearson's r(93) = .86, p=.000. For Version 2 and 3, r(03) = .83, p= .000.

[56] Person's r(93) = .30, p= .003. For Version 2 and 3, r(93) = .27, p= .009, and p= .008.

Testing Differences for Significance

A Wilcoxon signed rank test revealed that for some races and inference versions the difference to the self-reported demographics was insignificant, while in other instances the difference was significant. Test results are provided in Tables B1 and B2.

For Whites, the imputed race percentages of Version 2 were on average lower (median2=89.5%) than the true list percentages (median=92.5%). Versions 1 and 3 (using Hispanic-inclusive categories) were both on average higher (median1 = 93.5% and median3 =92.6%), but only Version 3 produced imputed race and ethnicity proportions that were insignificantly different from the true proportions. The imputed percentages for Blacks were indifferent from the true list percentages, regardless of the method, though each version of imputed Black percentages was on average higher than the true percentages. For Hispanics/Latinx, inferred proportions were on average higher than the true list proportions across versions. Yet, the first version derived from adults was the only version for which the differences between inferred and true list demographics were insignificant.

Table B1. Statistical Significance of the Differences between Self-Reported and Inferred Demographics

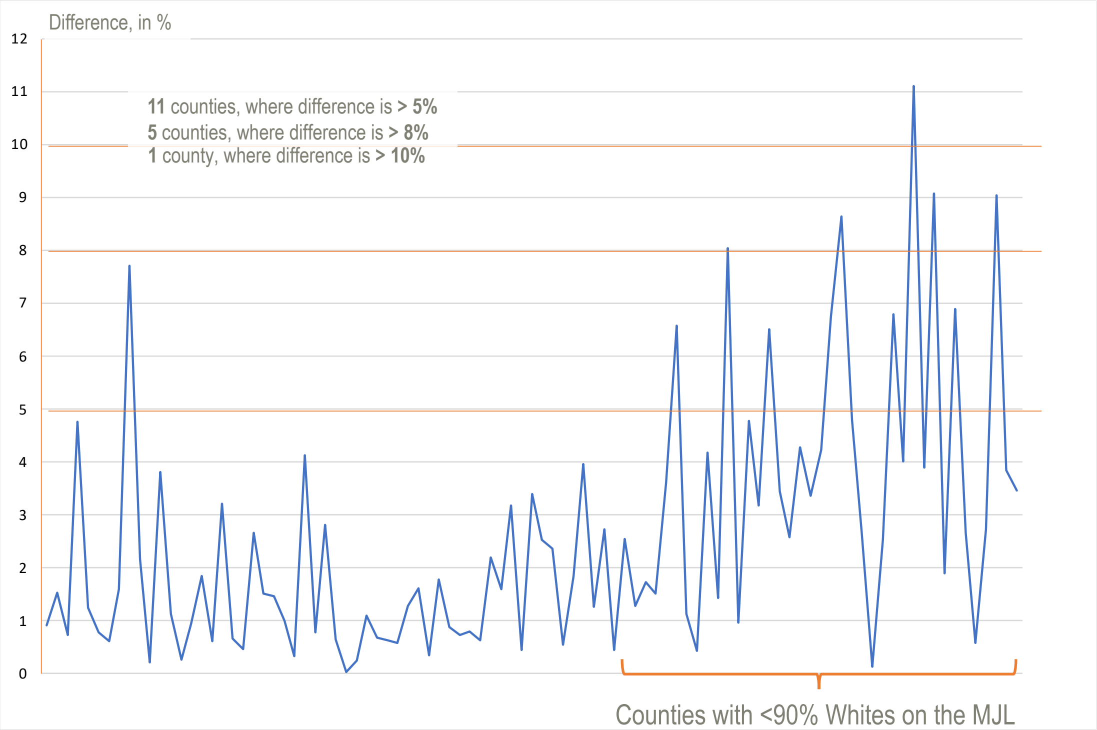
	White	Black	Native American	Asian	Hispanic/Latinx
Version 1	Significant	**Insignificant**	Significant	Significant	**Insignificant**
Version 2	Significant	**Insignificant**	Significant	Significant	Significant
Version 3	**Insignificant**	**Insignificant**	Significant	Significant	Significant

Table B2. Statistical Significance of the Differences between Self-Reported and Inferred Demographics (N=95 counties)

		White	Black	Native American	Asian	Hispanic/Latinx
Version 1	Ranked Sums z Score p	3396 -4.14 p = .000	2425 -0.54 p = .590	3480 -4.45 p = .000	3688 -5.23 p = .000	2628 -1.29 p = .196
Version 2	Ranked Sums z Score p	4393 -7.843 p = .000	2286 -0.022 p = .982	3034 -2.799 p = .005	3779 -5.56 p = .000	4083 -6.692 p = .000
Version 3	Ranked Sums z Score p	2697 -1.548 p = .122	2556 -1.024 p = .306	3667 -4.035 p = .000	3735 -5.401 p = .000	4083 -6.692 p = .000

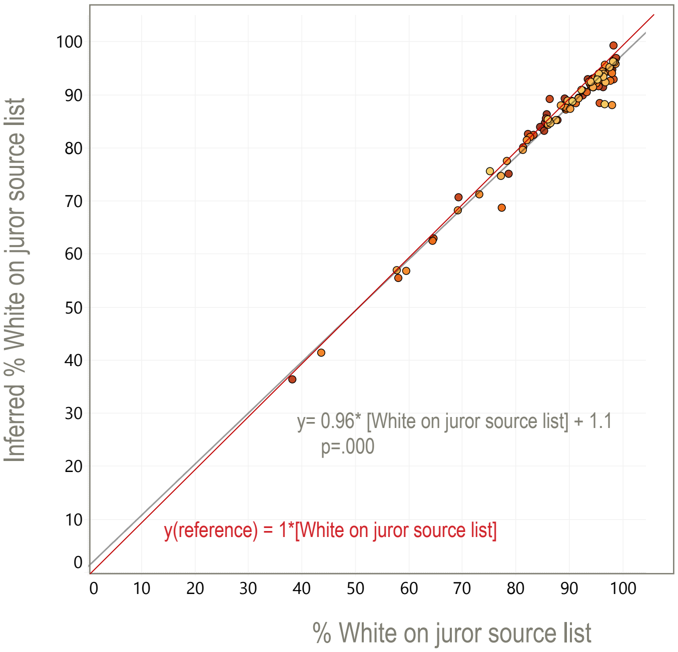
APPENDIX 2
Inference accuracy by county

Figure A. Absolute Differences between Inferred (Version 1) White % and True Master Jury List %, by County in the Order of White Population Size



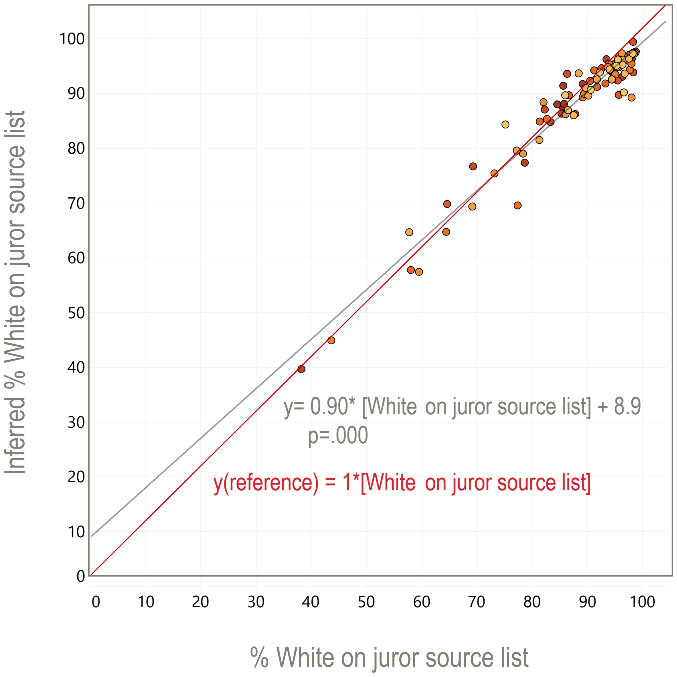
Figures B and C reveal the difference between using Hispanic-exclusive or -inclusive percentages for Whites when both are based on total populations. For counties with smaller White populations, as indicated in the scatterplot by a lower percentage on the master jury list (MJL), excluding Hispanics from the White count (Figure B) led to a closer estimate (though still statistically significant) than when including Hispanics (Figure C).

Figure B. White Proportions on the Juror Source List for Each County: Self-Reported Versus Inferred Demographics (Version 2)



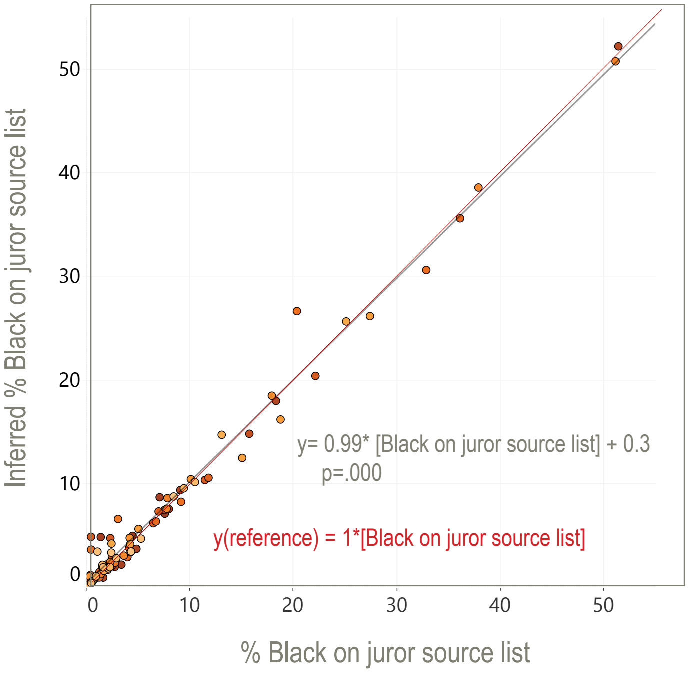
For counties with large White populations, including Hispanics led to a better estimate (as indicated by the proximity to the red line). Excluding them leads to underestimation of the White percentage on the list. Using adult-only, Hispanic-inclusive populations (Version 1) leads to a very similar, more accurate, estimate. Furthermore, the differences are statistically insignificant when Whites make up more than 90% on the master jury list.[57]

Figure C. White Proportions on the Juror Source List for Each County: Self-Reported Versus Inferred Demographics (Version 3)



Black inferred percentages are not significantly different from list percentages, and the size of the Black proportion does not impact the estimation error, as can be seen in Figure D. As shown, the fit line for inferred Black percentages nearly matches the reference line in the scatterplot, showing that, on average, the estimate is very close to the true list percentages and that this does not change with increasing or decreasing Black populations.

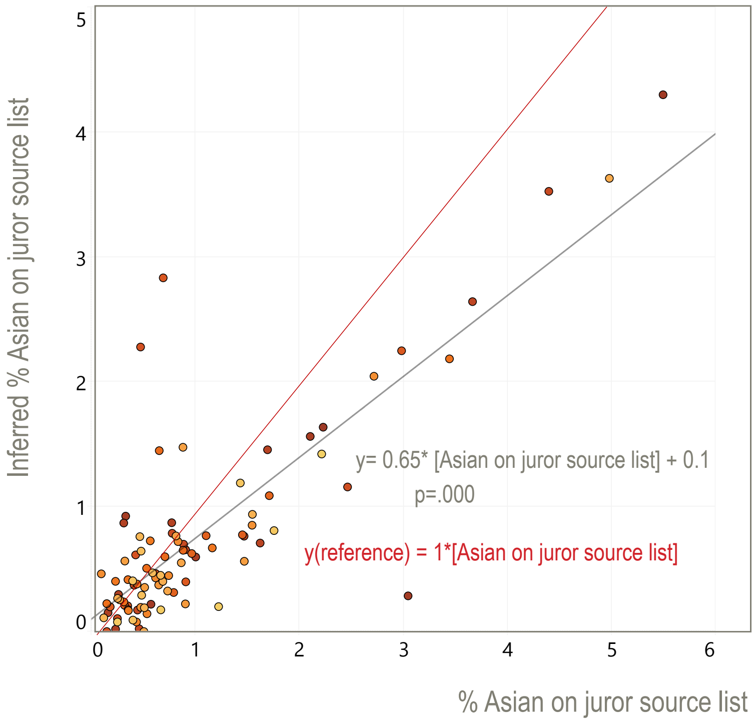
Figure D. Black Proportions on the Juror Source List for Each County: Self-Reported Versus Inferred Demographics (Version 3)



[57] A Wilcoxon signed rank test indicates that inferred White percentages in the remaining 56 counties were not statistically different from the list percentages (z= -.049, p = .961). However, this result does not hold for the other versions. The differences remain significant in Version 2 when splitting the sample between above and below 90% White populations. This is true also for Version 3, where the differences are only insignificant when including the complete sample.

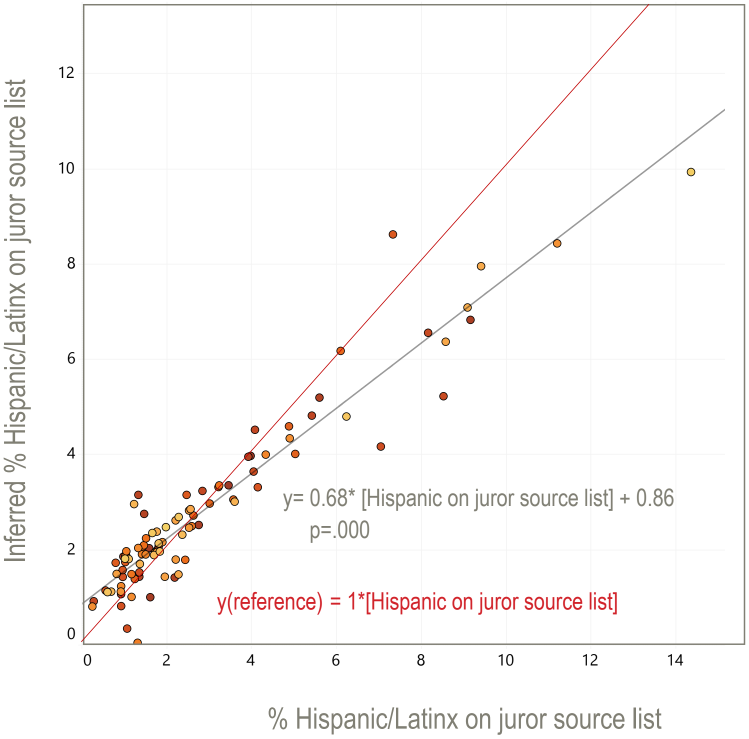
For Asians, where inference led to underestimated percentages in all versions, the scatterplot in Figure E reveals that the underestimation is greater for counties with larger Asian populations. However, the differences remain statistically significant even when excluding counties with above 1% Asians on the Master Jury List.

Figure E. Asian Proportions on the Juror Source List for Each County: Self-Reported Versus Inferred Demographics (Version 3)



In the case of Hispanic/Latinx percentages, using inference Version 1 as an example, where the estimation error was statistically insignificant, the size of a county's Hispanic population has a clear effect on the direction of the differences between inferred and true data. For counties with overall larger Latinx populations (above 3% on the master jury list, 29 counties), the true percentages were significantly higher than the inferred percentages. For counties with less than 3% Latinx on the list (66 counties), the true list percentages were on average significantly *lower* than the inferred percentages.[58] While Version 1 of inferred demographics may seem fairly accurate for Latinx overall, with only insignificant differences, the scatterplot in Figure F illustrates that caution may be warranted when using the results for individual counties. The clear overestimations for counties with smaller Hispanic/Latinx populations are balanced in this study by the underestimations in counties with higher Hispanic/Latinx populations.

Figure F. Hispanic/Latinx Proportions on the Juror Source List for Each County: Self-Reported Versus Inferred Demographics (Version 1)



[58] Both differences are statistically significant. For the sample above 3%, z= -3.56, p= .000. For the sample below 3%, z= -4.55, p= .000. Including children in the estimate does not seem to make a difference. Versions 2 and 3 show a very similar pattern, though somewhat less pronounced, significantly overestimating Hispanic percentages in small-population counties, and significantly underestimating in counties with high populations.

APPENDIX 3
Overall deviation from self-reported proportions

To summarize the distributional differences between the sets of inferred and those of the self-reported proportions, total deviation across the five estimated racial/ethnicity proportions were computed, each weighted by the racial group's self-reported proportion in the county.[59] The statewide averages and ranges of the overall weighted deviation for each county are listed in Table C below. Version 3 appears to be the most accurate overall, with a statewide weighted demographic proportion error of less than two percent.

Table C. Weighted Average Overall Deviation from Self-Reported Estimates

	Version 1 (Adults, inclusive race & ethnicity)	**Version 2** (Total Population, exclusive categories)	**Version 3** (Total Population, inclusive race & ethnicity)
Statewide Mean	2.28%	**2.11%**	1.95%
Statewide Range	8.99%	**9.49%**	8.44%

[59] See a similar approach working with dataset-wide proportions rather than county units in Marc N. Elliott et al., *A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity*, 43 Health Serv. Res. 1722–1736 (2008).

APPENDIX 4
Disparity Results by Inference Version

In the comparison of absolute disparity measures shown below, negative values reflect underrepresentation; positive values reflect overrepresentation.

Table D. Absolute Disparities (95 Counties), in %

		Mean %	Standard Deviation %	Median %	IQR %	Statistical Significance of Disparity Deviation[60]
White	Self-Reported	-1.8	3.8	-103	4.2	
	Estimate 1	-0.4	1.1	-0.4	1.1	underestimation significant
	Estimate 2	-3.9	2.7	-3.4	3.2	overestimation significant
	Estimate 3	-1.4	1.4	-1.2	1.7	underestimation insignificant
Black	Self-Reported	-0.4	1.7	0	0.9	
	Estimate 1	-0.2	0.9	0	0.4	accurate
	Estimate 2	-0.2	0.9	-0.1	0.5	slight overestimation insignificant
	Estimate 3	-0.2	0.9	-0.1	0.5	slight overestimation insignificant
Native	Self-Reported	-0.1	0.3	-0.1	0.3	
	Estimate 1	0	0.1	0	0.1	underestimation significant
	Estimate 2	0	0.2	0	0.1	underestimation significant
	Estimate 3	0	0.2	0	0.1	underestimation significant
Asian	Self-Reported	0.5	0.7	0.4	0.6	
	Estimate 1	0.3	0.3	0.2	0.3	underestimation significant
	Estimate 2	0.3	0.4	0.2	0.3	underestimation significant
	Estimate 3	0.3	0.4	0.2	0.3	underestimation significant
Hispanic/ Latinx	Self-Reported	1.2	2	0.7	1.8	
	Estimate 1	1.2	1.1	0.8	0.3	slight overestimation insignificant
	Estimate 2	1.9	1.7	1.5	1.6	overestimation significant
	Estimate 3	1.9	1.7	1.5	1.6	overestimation significant

Overall, deriving disparities from Hispanic-inclusive, adult populations (Version 1), we would have falsely assumed a more representative master jury list for Whites, Native Americans, and Asians. For these, the deviations from the true disparities are statistically significant. Nevertheless, the true Hispanic/Latinx overrepresentation is very closely estimated, only insignificantly overestimating this disparity.

Excluding Hispanics from the inferred race proportions (Version 2) leads to statistically significant overestimation of White and Native American underrepresentation and to significant underestimation of Asian overrepresentation. Even when including Hispanics in the racial count but using total populations (Version 3), Asian and Native American disparities are still underestimated. With this version, Hispanics would have seemed significantly more overrepresented on the list than they are, but the slight underestimation of White underrepresentation is statistically insignificant. The true underrepresentation of Blacks on the list is accurately shown by the disparity measures derived from inferred demographics, regardless of the method, and any deviation is insignificant.

[60] Statistical significance was measured with a Wilcoxon Signed Rank Test. Over-and under-estimation of the disparity measures are here listed as statistically insignificant when $p > .1$. Statistically significant measures here were at least significant at the $p<.001$ level.

Similarly, comparative disparity measures derived from inferred demographics can be evaluated against the measures derived from self-reported data and tested again for statistical significance of the differences to the true disparities. The results mirror those presented for absolute disparities in every cell and are shown in Table E.

Table E. Comparative Disparities

		Mean %	Standard Deviation %	Median %	IQR %	Statistical Significance of Disparity Deviation[60]
White (counties: 95)	Self-Reported	-2.2	4.6	-1.4	4.9	/
	Estimate 1	-0.5	1.3	-0.5	1.2	underestimation significant
	Estimate 2	-4.5	3.5	-3.7	3.7	overestimation significant
	Estimate 3	-1.6	1.8	-1.3	2.0	slight underestimation insignificant
Black (counties: 62*)	Self-Reported	-0.6	2.0	0.0	1.5	/
	Estimate 1	-1.5	18.1	-1.3	8.1	overestimation insignificant
	Estimate 2	-3.8	18.8	-2.6	11.4	overestimation insignificant
	Estimate 3	-1.7	20.2	-1.7	9.7	overestimation insignificant
Native (counties: 1*)	Self-Reported	92.8	/	-92.8	/	/
	Estimate 1	-58.4	/	-58.4	/	underestimation significant
	Estimate 2	-66.8	/	-66.8	/	underestimation significant
	Estimate 3	-66.8	/	-66.8	/	underestimation significant
Asian (counties: 3*)	Self-Reported	120.2	29.2	134.4	/	/
	Estimate 1	66.5	15.1	69.8	/	underestimation significant
	Estimate 2	68.4	19.9	70.5	/	underestimation significant
	Estimate 3	69.4	19.0	70.9	/	underestimation significant
Hispanic/ Latinx (counties: 25*)	Self-Reported	101.1	84.5	96.5	111.2	/
	Estimate 1	75.8	39.5	74.8	53.7	underestimation insignificant
	Estimate 2	121.4	58.5	107.3	82.9	overestimation significant
	Estimate 3	121.4	58.5	107.3	82.9	overestimation significant

* only including counties with eligible populations >2%

[61] Statistical significance was measured with a Wilcoxon Signed Rank Test and is based on the full sample of 95 counties. Over- and underestimation of the disparity measures are here listed as statistically insignificant when p >.05.

To assess the disparity inaccuracies across races in each version, the deviations from the true disparities across races are averaged, each weighted by the eligible population in each county. As was concluded for the overall deviations of the demographic estimates, Version 3 provides estimated disparities that show the smallest state-wide overall deviation from the true distribution of disparities. State-wide averages of the weighted overall deviations for absolute disparities are shown in Table F.

Table F. Weighted Average Overall Deviation from Absolute Disparities Derived from Self-Reported Estimates

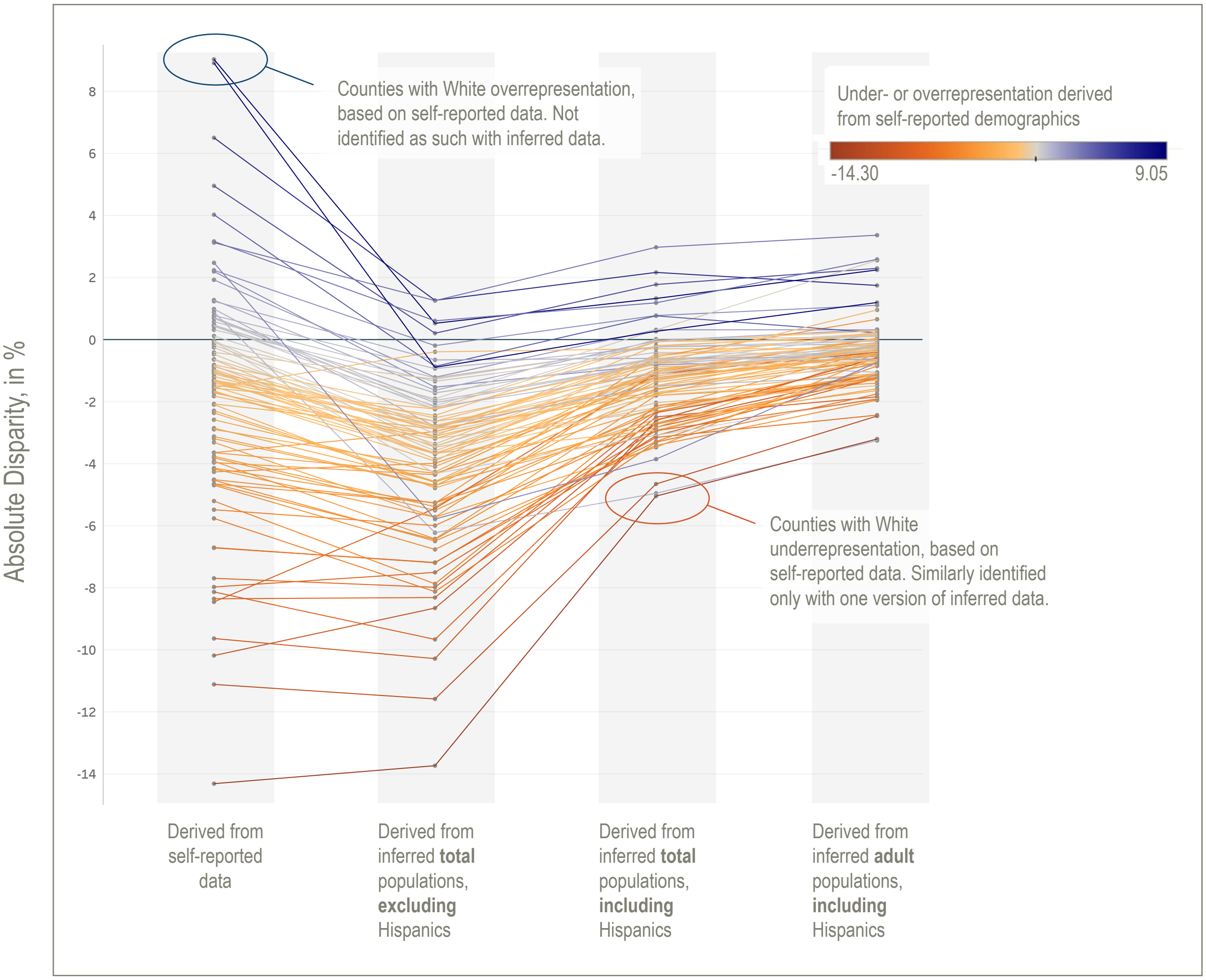
	Version 1 (Adults, inclusive race & ethnicity)	**Version 2** (Total Population, exclusive categories)	**Version 3** (Total Population, inclusive race & ethnicity)
Statewide Mean	2.35%	**2.10%**	1.99%

APPENDIX 5
Disparity Results by County

Figure G shows the distribution of absolute disparities in comparison between versions for Whites. The inference-derived disparities have a much smaller range across counties than the true disparities do, but this is especially evident for the methods including Hispanics in the racial count (Versions 1 and 3).

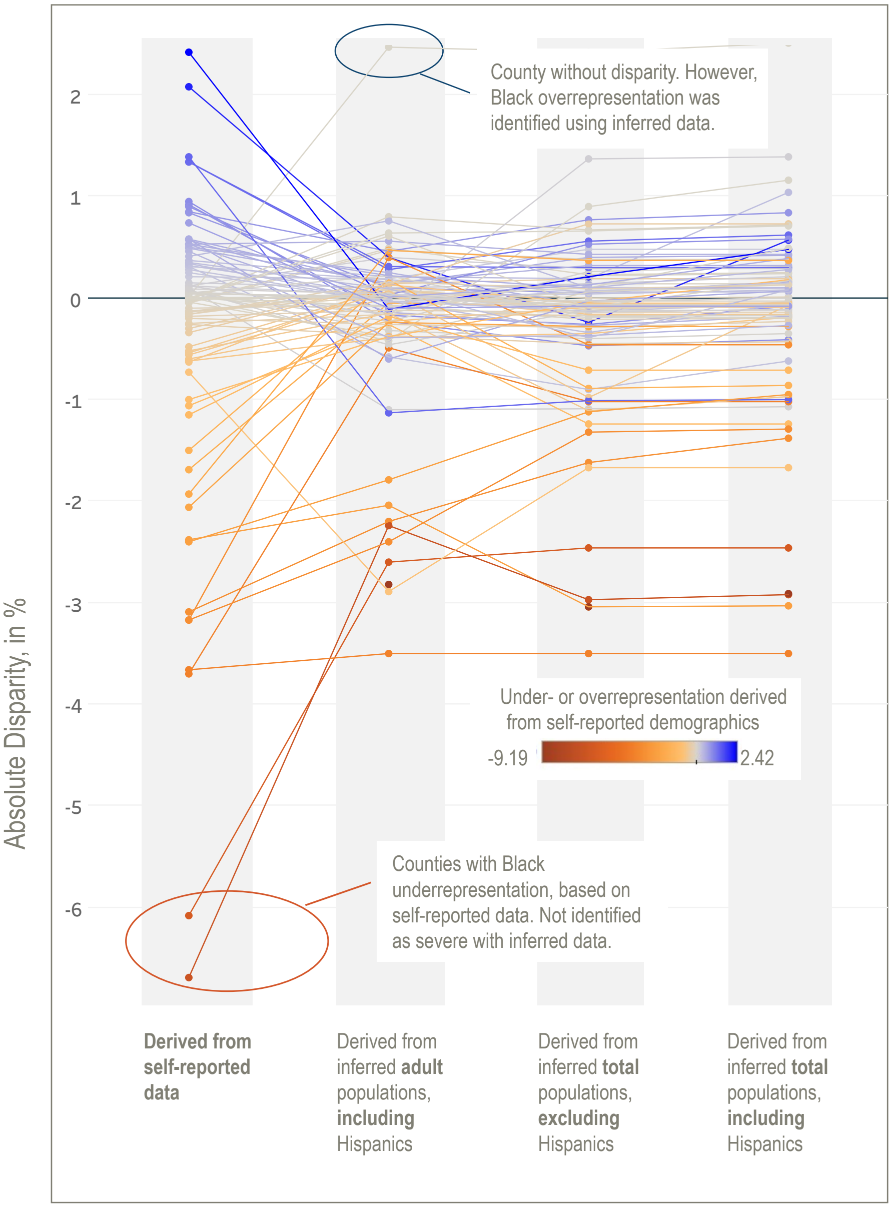
While Version 3 was the version with overall statistically insignificant deviations to true White proportions and disparities and statewide averages closest to the true averages, this figure illustrates that there are some counties where the true extent of underrepresentation is only picked up when using total populations and Hispanic-exclusive categories for imputation. Furthermore, none of the versions can reproduce the original absolute disparity findings for the few counties with White overrepresentation above 3% absolute disparity.

Figure G. Disparities for Whites Between Juror Source List and County Population in 95 Counties, by Data Source



For Black demographics, which appeared to be successfully inferable, Figure H depicts a similarly smaller range of disparities when imputing race. This might be especially problematic for the few outlier counties where the true disparities are at concerning levels of underrepresentation, but the same counties would not show similarly high disparities when imputing race.

Figure H. Disparities for Blacks Between Juror Source List and County Population in 95 Counties, by Data Source



County without disparity. However, Black overrepresentation was identified using inferred data.

Counties with Black underrepresentation, based on self-reported data. Not identified as severe with inferred data.

APPENDIX 6
Distribution of Disparity Results by Range

Table F shows the distribution of disparities depending on the source data used for Whites, Blacks, and Hispanics. For Whites, where the original analysis identified 3 out of 95 counties with more than 10% absolute disparities, indicating White underrepresentation, only Version 2 led to the identification of the same three counties.

Table F. Comparison of Representativeness Results for Tennessee Counties, Based on Each Version of Demographics

	Disparities based on self-reported demographics	…based on imputed demographics (Version 1)	…based on imputed demographics (Version 2)	…based on imputed demographics (Version 3)
Absolute Disparities for Whites				
Less than -10%	-3		3	
-10% to -5%	11		24	1
-5% to 5%	78	95	68	94
5% to 10%	3			
Absolute Disparities for Black				
-10% to -5%	3			
-5% to 5%	92	95	95	95
Absolute Disparities for Hispanic/Latinx				
-5% to 5%	89	94	88	88
5% to 10%	6	1	7 (1 new)[62]	7 (1 new)
Comparative Disparities for Whites (N=95)				
-20% to 20%	95	95	95	95
Comparative Disparities for Black (N=62*) (and # of flagged counties added when disparities were based on estimates)				
Less than -50%	6 (5 not flagged in estimates)	1	1	1
-50% to -40%	1		1 (new)	1 (new)
-40% to -20%	3	2 (new)	5 (3 new)	4 (2 new)
-20% to 20%	48	58	54	54
20% to 40%	3			
40% to 50%				
More than 50%	1	1 (new)	1 (new)	2 (new)
Comparative Disparities for Hispanic/Latinx (N=25*) (# of flagged counties added, or disparity direction changes compared to true disparity)				
-40% to -20%	2			
-20% to 20%	2	2		
20% to 40%	1	2 (1 reversed direction)	2 (both reversed direction)	2 (both reversed direction)
40% to 50%	1	3		
More than 50%	19 (2 not be flagged as high in estimates)	18 (1 reversed direction)	23 (2 new, 2 reversed direction)	23 (2 new, 2 reversed direction)

* Only including counties with populations of this race/ethnicity >2%

[62] This indicates that one of the seven counties within this range was not identified in the same range in the original analysis.

For Blacks, the absolute disparities calculated based on inferred demographics are smaller than 5% for all counties, whereas the original analysis had identified three counties with underrepresentation between 5% and 10% absolute disparity. Furthermore, the original analysis had flagged six counties where Blacks were clearly underrepresented on the jury list with above 50% comparative disparities (CD), which is often a better measure due to the small overall Black population. Only one out of these six would have been identified when using estimated racial counts. The one county flagged for Black *overrepresentation* with above 50% CD is not identified at all via the estimated versions. What is more, the disparity analyses based on inferred race percentages identify one to two new counties each with seemingly great overrepresentation above the threshold, which were not marked as such in the original analysis. Similarly, the somewhat less severe underrepresentation ranges of 20% to 40% and 40% to 50% CD included four counties in the original analysis, only two in estimated Version 1, but six in Version 2, and five in Version 3. Yet, none of the original counties within these ranges are found in the same ranges in Version 1, and only half are identified in these ranges in Versions 2 and 3. Instead, other counties are identified with underrepresentation in these ranges.

For Latinx, who have even smaller county populations than Blacks, the absolute disparities were not above 10% in either analysis. Inference-based Versions 2 and 3 of the analysis identified the same six counties that the original analysis had identified as between 5% and 10% overrepresentation. However, most of the counties with populations above 2% (high enough to compute comparative disparities) had very high comparative disparities, indicating great Hispanic/Latinx overrepresentation on the list. The two counties that the original analysis had identified with *underrepresentation* of 20% to 40% are not identified in any of the other analyses. Those with disparities above the threshold of 50% for overrepresentation, which are the majority of the 25 counties for which comparative disparities can be calculated, are, however, indeed identified as such, or nearly as such, in each one of the inference-based analyses. Each of these analyses, nonetheless, flagged at least one additional county with apparently severe Hispanic/Latinx overrepresentation, while, according to the original analysis, Hispanics/Latinx were even underrepresented on the list in these specific counties. Thus, for these few counties, using inferred ethnicity percentages led to results that were opposite to the original findings, reversing the direction from minor underrepresentation to severe overrepresentation.

State Justice Institute

This white paper was written with support from the State Justice Institute
under grant number SJI-22P006. The points of view expressed are those of the author
and do not necessarily represent the official position or policies of the State Justice Institute.

ncsc.org