How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved

Lingwei Cheng*
Carnegie Mellon University
Pittsburgh, PA, USA

Isabel O. Gallegos*
Stanford University
Stanford, CA, USA

Derek Ouyang
Stanford University
Stanford, CA, USA

Jacob Goldin
University of Chicago
Chicago, IL, USA

Daniel E. Ho
Stanford University
Stanford, CA, USA

ABSTRACT

We address two emerging concerns in algorithmic fairness: (i) redundant encodings of race – the notion that machine learning models encode race with probability nearing one as the feature set grows – which is widely noted in theory, with little empirical evidence; and (ii) the lack of race and ethnicity data in many domains, where state-of-the-art remains (Naive) Bayesian Improved Surname Geocoding (BISG) that relies on name and geographic information. We leverage a novel and highly granular dataset of over 7.7 million patients' electronic health records to provide one of the first empirical studies of redundant encodings in a realistic health care setting and examine the ability to assess health care disparities when race may be missing. First, we show that machine learning (random forest) applied to name and geographic information can improve on BISG, driven primarily by better performance in identifying minority groups. Second, contrary to theoretical concerns about redundant encodings as undercutting anti-discrimination law's anti-classification principle, additional electronic health information provides little marginal information about race and ethnicity: race still remains measured with substantial noise. Third, we show how machine learning can enable the disaggregation of racial categories, responding to longstanding critiques of the government race reporting standard. Fourth, we show that an increasing feature set can differentially impact performance on majority and minority groups. Our findings address important questions for fairness in machine learning and algorithmic decision-making, enabling the assessment of disparities, tempering concerns about redundant encodings in one important setting, and demonstrating how bigger data can shape the accuracy of race imputations in nuanced ways.

ACM Reference Format:
Lingwei Cheng, Isabel O. Gallegos, Derek Ouyang, Jacob Goldin, and Daniel E. Ho. 2023. How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved. In *2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23), June 12–15, 2023, Chicago, IL, USA.* ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3593013.3594034

*Equal first author



This work is licensed under a Creative Commons Attribution International 4.0 License.

FAccT '23, June 12–15, 2023, Chicago, IL, USA
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0192-4/23/06.
https://doi.org/10.1145/3593013.3594034

1 INTRODUCTION

Racial equity assessment is critical for understanding disparate impacts of policies on protected groups.[1] U.S. Executive Order 13,985 [29], for instance, requires federal agencies to conduct racial equity assessments in order to reduce, eliminate, and prevent racial discrimination and inequities. A core challenge in conducting such assessments, however, is often the lack of recorded or reported race. Due in part to the Privacy Act of 1974 [38], many agencies do not possess race information. Similarly, in consumer finance, for instance, despite the U.S. Consumer Financial Protection Bureau's mandates to prohibit disparities along race, many lenders either do not collect or are federally prohibited from collecting information about race [13, 17].

Conducting racial disparity assessment when race is itself not directly observed has become an increasingly important and relevant methodological question, given the ubiquitous use of data-driven decision-making tools. Analyzing only complete cases with no missing race may underestimate the true level of disparity [57]. To address this, researchers have proposed a number of techniques and methods to bypass the direct collection of race, with imputation being an important alternative. The current state-of-the-art imputation approach is Bayesian Improved Surname Geocoding (BISG), which uses demographic information about census geographies (*e.g.*, census tract, census block group) and surnames to impute race [28, 47]. Bayesian Improved First Name Surname Geocoding (BIFSG) additionally incorporates first name-based probabilities [95]. However, using either imputation method to estimate disparity is known to be subject to strong assumptions [15, 51].

As imputation relies on features correlated with race, it raises an important concern of "redundant encodings" which is central to the discussions surrounding privacy and algorithmic fairness [3, 9, 19, 20, 27, 42, 54, 61, 77, 91]. The leading fairness and machine learning (ML) textbook, for instance, considers a set of independent features Z, each of which is slightly correlated with a sensitive attribute such as race R [9]. As the size of Z grows, the probability that a classifier can infer the sensitive attribute $P(R|Z)$ increases and can approach one.[2] This conception of redundant encodings is of particular legal

[1] For brevity, we use "race" to refer to both race and ethnicity throughout the paper.
[2] Redundant encodings can also be defined in the context of an equivalent classifier; in this definition, a feature set Z redundantly encodes a sensitive attribute R if a classifier trained on a feature set that includes the sensitive attribute, namely $P(Y|Z,R)$, is equivalent to a classifier trained on equivalent features with the sensitive attribute

consequence, as substantial parts of U.S. anti-discrimination law rest on a principle of "anti-classification," whereby state or covered entities should not classify, or make decisions, based on race [44].[3] If an algorithm does not include race in its feature set, which naively meets the demands of anti-classification, but is able to recover race from the remaining features that are highly correlated with race, it risks violating anti-discrimination law by functionally engaging in disparate treatment. The notion that the bigger the data, the bigger the risk of redundantly encoding race features prominently in the legal, policy, and algorithmic fairness canon [3, 9, 19, 20, 27, 42, 54, 61, 77, 91]. While Barocas et al. [9] and Hardt [42] make compelling theoretical cases for such concerns, the textbook articulation relies on strong assumptions – most notably independence of features Z – and surprisingly little is known about redundant encodings in the wild (*i.e.*, large-scale administrative datasets beyond benchmark or toy datasets).

In this paper, we leverage a unique and novel dataset of electronic health records (EHR) on over 7.7 million patients to understand what happens when we move beyond conventional BI(F)SG methods by (a) using supervised ML to impute race for disparity assessments, and (b) incorporating rich features beyond name and geography. We ask the following research questions: (i) How does an ML-based approach improve imputation of race compared to the BI(F)SG baseline? (ii) Does using richer features heighten concerns of redundant encodings or, put differently: how redundant are redundant encodings in the wild? (iii) Can we use an ML-based approach to disaggregate broad racial categories and make more precise estimates for subgroups? (iv) Do these ML methods coupled with the use of rich feature sets perform equally well for different racial groups?

Our study makes four important contributions to the empirical investigation on disparity assessment and redundant encodings. First, we show that supervised learning with BIFSG inputs leads to performance improvements, driven by improved imputations for minority groups. The intuition here is that supervised learning can identify base rate differences, interactions, and non-linearities in BIFSG prior information (about demographic correlates of names and census geographies) compared to conventional methods that rely on strong independence assumptions. Non-linearities, in particular, appear to drive gains from ML. Second, we assess the gains from incorporating patient information beyond BIFSG features. By using over 1,000 features of patients including insurance plans, allergies, vital signs, diagnoses, medical procedures, and more, we show that such information in fact provides little marginal information about patient race beyond the calibrated BIFSG model. The benefits of the ML approach over conventional BIFSG arise primarily from model complexity instead of additional features. This finding is surprising and calls into question conventional wisdom about big data and redundant encodings in one important health care setting. Put differently, bigger data does not appear to drive redundant encodings in this setting. Third, we demonstrate these findings with a disaggregation of the Asian and Pacific Islander (API) category into (i) Asian and (ii) Native Hawaiian or Other Pacific Islander (NHPI) groups. This disaggregation addresses longstanding critiques of legacy standards of government race reporting [12, 32, 52] that employ only five racial categories, and enables a more nuanced assessment of health care disparities. Fourth, we show that expanding the feature set itself can have varying impacts on subgroup prediction performance, and results in different precision and recall trade-offs for each racial group depending on the category of the data and its quality.

At the outset, we note that race is socially constructed and does not fit neatly into the conventional discrete measure assumed in many algorithmic fairness papers [41]. We discuss the ethical concerns and social impact of our work in Appendix A. Our approach is meant to inform settings where racial disparities cannot be assessed because of the lack of self-reported race information, and we illustrate how to expand subgroup analysis in this setting. Redundant encodings raise challenging trade-offs. If race is easy to infer, as the textbook definition of redundant encodings suggests, our ability to detect racial disparities is very strong – but so is the threat of big data-driven algorithms engaging in potentially discriminatory or illegal behavior. On the other hand, if race is difficult to infer, concerns of such algorithms violating anti-discrimination law may be moderated, but our ability to assess disparities would be similarly diminished. Understanding these trade-offs with "awareness" is critically important for advancing fairness in ML and how anti-discrimination law handles algorithmic decision-making.

2 RELATED WORK

The risk of redundant encodings is a central concern in algorithmic fairness [3, 9, 19, 20, 27, 42, 54, 61, 77, 91]. The increasing size of data used in ML has raised new legal concerns of proxy discrimination. AI systems may learn a protected attribute that is not explicitly measured but is encoded in proxies that are correlated with the protected attribute [4, 77]. Merely removing protected attributes from the data thus does not guarantee non-discrimination [20, 75] and may even hurt the protected group [19, 27]. An often cited example of proxy discrimination is the use of ZIP Code in the illegal practice of redlining. A recent study [36] uses rich Boston Federal Home Mortgage Disclosure Act data containing information on mortgage applications and demonstrates a greater ability to predict race using traditional credit pricing inputs, compared to using ZIP Codes. This suggests that common intuitions about which variables govern as race proxies might be misleading. Despite the strong ties between socioeconomic status, health, and race, there are very few studies that systematically investigate the claim of redundant encodings and show the extent to which these factors are related to race in a realistic big data setting. Indeed, precisely because such variables are highly correlated with one another, the textbook invocation of redundant encodings (which assumes independence across features) may not generalize to all settings.

Despite extensive work on ML in health applications, less is known about redundant encodings in the medical setting. Some emerging work has illustrated the potential for health data to be correlated with race. Gichoya et al. [35] shows that computer vision models with medical imagery can predict patients' self-reported

removed, namely $P(Y|Z)$. Because the equivalent classifier condition can be trivially satisfied if R is irrelevant to Y or if R is fully mediated through Z, we focus on the first definition of redundant encodings.

[3] A common alternative conception of anti-discrimination law is anti-subordination, which posits that the law should help groups that have been historically subordinated, and racial classifications may hence be countenanced [84].

race with an area under the receiver operating characteristic curve (AUROC) range as high as 0.91–0.99 for X-ray images. Duffy et al. [26] finds that deep learning models with cardiac ultrasound imagery are able to identify age and sex, but unable to reliably predict race, and that predictions of race are associated with tuning the proportion of confounding variables such as age or sex. In a study of Boston and New York City patients, Adam et al. [1] finds that race information can be subtly embedded in clinical notes, with models distinguishing between White and Black patients with an AUROC as high as 0.83, while over 40 physicians were unable to identify patients' race from the same notes. While existing work is suggestive, it does not assess the marginal predictive power of medical features to formally assess concerns about redundant encodings. This is particularly important given existing BI(F)SG baselines. Moreover, existing literature has focused on a limited set of racial categories (*e.g.*, Black and White), but conventional measures of performance (*i.e.*, AUROC) may not be appropriate in the presence of larger class imbalances.

Our work also pertains to the literature on the use of BI(F)SG – the dominant method to assess disparities when race is missing. This approach has been used extensively in a variety of fields including finance [18, 56], elections [23], and health [24]. BIFSG's incorporation of first name priors further improves imputation accuracy [95]. Despite the popularity of BI(F)SG, recent work has documented the limitations of BI(F)SG and related proxy approaches for assessing disparities [*e.g.*, 15, 103]. BI(F)SG also makes strong conditional independence assumptions, namely $P(G|R,S) = P(G|R)$ and $P(F|R,S,G) = P(F|R)$ for race R, first name F, surname S, and geographic area G, which are often violated in practice. Additionally, BI(F)SG is limited by Census measurement errors: minority groups may be under-counted in census blocks, and Census name tables contain only the most frequent names, which disproportionately exclude names more common in minority groups [48]. Validation studies have found that BI(F)SG performs best for White, Hispanic, and API populations, but has lower accuracy for Black and American Indian and Alaska Native (AIAN) populations and women [2, 24]. Lastly, BI(F)SG relies on name priors for the combined API category which has become outdated since the Office of Management and Budget (OMB) revised the standards in 1997 [72] to disaggregate the category into Asian and NHPI. In this study, we demonstrate how ML can address some of these limitations of BI(F)SG and related disparity assessments.

To overcome shortcomings of BI(F)SG, several studies have proposed alternative imputation methods, with some including additional features [40, 102], and others improving on model flexibility using ML [22, 55, 64, 99]. Matthews et al. [64] and Xue et al. [99] show that using ML models with BISG and additional demographic features improve upon BISG when evaluated on AUROC. However, the work is limited to simplified racial categories [64, 99] or lacks a meaningful BI(F)SG baseline for comparison [99]. Decter-Frain [22] shows that ML methods provide better-calibrated imputations for individual racial groups, particularly for Asian and Hispanic individuals in the voting setting of Florida, Georgia, North Carolina, and California. Kim et al. [55] trains a multilayer perceptron neural network on 15,000 demographic and diagnosis code features from a dataset of over 1.5 million patients, and finds it outperforms other less complex supervised learning algorithms; it is limited, however, by only considering White, Black, Hispanic, and Other racial groups, and exclusively studying Chicago and New York City patients. These studies suggest that more data and more flexible models can improve imputation performance, but offer fewer insights into *how* the additional flexibility and richer features impact race imputation or how they are connected to the problem of redundant encodings. There is also often more personal information in the health care domain as compared to in voting records.

Last, our work speaks to existing work critiquing conventional race categories as aggregating individuals who come from greatly varying cultural and socioeconomic backgrounds [8, 50, 82, 90]. The API category, for instance, is widely used but can mask health disparities between vastly divergent sub-populations [6, 37, 46, 86]. In health care, understanding these contexts is central to studies of the social determinants of health, health care access, and health outcomes. As such, numerous researchers and policymakers have advocated for further disaggregation of race categories to better understand health inequities [52, 81, 100]. OMB's Chief Statistician has proposed updating the 1997 reporting standards to improve disaggregation [74]. Given the significantly different results we uncover for racial inference for Asian versus NHPI populations, our study provides a demonstration of the importance of disaggregation for a more complete understanding of racial disparities and offers a potential way forward.

3 METHODOLOGY

3.1 Data Source and Study Population

We use a unique dataset to study the question of race, health disparities, and redundant encodings in a consequential setting. The American Family Cohort (AFC) [94] contains electronic health care records for over 7.7 million U.S. patients from 2010 to the present, including patient names, addresses, self-reported demographic information, insurance, allergies, and medical diagnostic codes. The data is derived from the American Board of Family Medicine PRIME Registry [79], which is currently the largest national Qualified Clinical Data Registry for primary care, with AFC practices in 47 states and patients from all 50 states. Primary care clinicians opt in to this system which helps them improve patient outcomes and alleviate the burden of reporting quality measures for value-based payment models. The AFC dataset contains populations that are often underserved and missing from other medical data sources, including rural, low-income, and racial minority populations. It also includes patients on private insurance plans as well as on Medicaid and Medicare. This dataset is larger and more geographically-diverse than urban-focused or hospital-based ones used in prior race imputation and redundant encoding works, with a higher volume and more diverse cross-section of patients presenting to primary care [39, 49, 97]. Moreover, this dataset is uniquely suited for our research aims, as it contains first and last name information and residential information to construct the BI(F)SG baseline, self-reported race to assess performance of imputations (which is often missing), and highly granular health information in a domain where ML is increasingly deployed.

We include patients who satisfy the following three criteria. First, we include patients living in the 50 states or Washington, D.C. We exclude patients from other overseas U.S. territories or military

bases. Second, we require patients to have at least one record dated between 2010 and 2022 and a recorded surname. This ensures that we have enough data and at least last name information to use BI(F)SG imputation (where the algorithm falls back to BISG when first name is unavailable) as our baseline for comparison. Third, in order to assess performance, we require self-reported race and ethnicity (AIAN, Asian, Black or African American (Black), Hispanic or Latino (Hispanic), NHPI, or White). For cleaning and coding of self-reported fields, see Appendix B. Because the vast majority of health care providers do not use multiracial identification in our data, we exclude patients assigned to multiple races, which represents less than 0.05% of the total sample.[4] These inclusion criteria allow us to maintain a study sample that is drawn from the U.S. general population and contains sufficient information for the imputation tasks and comparison. Our final sample includes 5,178,620 unique individuals. Table 1 summarizes the distribution of the study population's race and ethnicity.

All work was conducted on servers approved for High Risk and Protected Health Information (PHI) data, and the research was approved by the Institutional Review Board.

3.2 Features

For comparing the BIFSG and ML imputation methods, we construct the target label to use the minimum acceptable combined race and ethnicity categories outlined in the 1977 OMB standards, namely AIAN, API, Black, Hispanic, and White.[5] In our first analyses, we combine Asian and NHPI to the API category to (a) assess performance relative to the minimum OMB standard, and (b) track conventional BI(F)SG practice, which relies on name tables published by the Census Bureau and under the Home Mortgage Disclosure Act that collapse Asian and NHPI into a single category [14, 92, 93]. One benefit of the AFC dataset is that it enables us to explore how ML with the AFC dataset can facilitate disaggregation of Asian and NHPI groups, which OMB adopted in 1997.

We derive a rich set of features from the AFC dataset, which are presented in Table 1. The sets of features are ranked by their accessibility to researchers, with the baseline features being the most likely to be available and the medical features being the least likely to be available to researchers facing administrative data where race is missing.

Baseline features comprise the prior probabilities used with the conventional BIFSG imputation method, namely first name $P(F|R)$, surname $P(R|S)$, and geographic area $P(G|R)$ priors, for race R, first name F, surname S, and geographic area G.[6] We obtain $P(R|S)$ from the 2010 Decennial Census Surname Files [14, 93], $P(F|R)$ from the Home Mortgage Disclosure Act Loan Application Registers [92], and $P(G|R)$ from the American Community Survey 2014-2018 5-Year Summary at the census block group (CBG) and state levels. As examples, features for a person named "Michael Smith" or "Jose Lopez" in two different CBGs are shown in Table 2.[7] Following [95], we revert to BISG when only surname is available.

Demographic features comprise age, gender, marital status, and patients' state of residency. State is treated as a demographic feature distinct from geographic priors in BIFSG, as it is an indicator for the person's residency, not a probabilistic prior. Gender, marital status, and state are one-hot encoded, and age at the time of the most recent visit is treated as a continuous value. *Behavioral features* intend to capture patients' history with substance use and their social and psychological characteristics, including smoking history and language preference; all behavioral features are one-hot encoded. *Administrative features* intend to capture how patients interact with the health care system, including insurance plans, which are one-hot encoded, and the frequency of physician visits for each year from 2010 to 2022, which are continuous values. *General health features* include allergies, immunizations, and vital signs. Allergies are one-hot encoded, while the count of immunizations is treated as a continuous value. Vital signs, including height, weight, systolic blood pressure, diastolic blood pressure, body mass index, heart rate, respiratory rate, oxygen saturation, and temperature, are also treated as continuous values and aggregated across all available records for each patient as a linearly weighted average.[8]

Medical features comprise specific diagnoses and procedures stemming from visits. We take advantage of the substantive hierarchical structure of diagnosis and procedure codes to categorize these codes and avoid sparsity. The diagnosis codes are available in International Classification of Diseases Ninth Revision (ICD-9) and Tenth Revision (ICD-10) and SNOMED Clinical Terms (CT) formats, each of which are internationally transferable identification codes for medical diagnoses. ICD-9 codes and SNOMED CT codes are mapped to their ICD-10 equivalent and one-hot encoded. One of the challenges of working with diagnostic codes is high dimensionality. For instance, there are over 68,000 ICD-10 codes [10]. One-hot encoding all of these, when coupled with the size of our dataset, make computation and hyperparameter searching challenging for tree-based classification approaches (*e.g.*, random forest). We hence preprocess and select eligible diagnostic codes using two heuristics. First, we conduct a substantive literature review to compile all diagnostic codes that are known to present health disparities across race. Appendix D shows a list of such studies and the corresponding ICD codes. Second, we select features that are empirically correlated with race. We perform a χ^2 test and calculate the mutual information between diagnosis codes and race class labels on the training data; diagnosis codes with the highest mutual information and $p \leq 0.05$ are selected. These heuristics perform quite well based on a qualitative understanding of diagnostic codes that have the potential to encode race. To process procedure codes, Current Procedural Terminology (CPT) codes are grouped at the Category I and II levels, Healthcare Common Procedure Coding System (HCPCS) codes are grouped at the Level II level, and the

[4] Although one of our major efforts is to disaggregate racial categories, the multi-racial sample size is too small to reliably capture the diversity and complexity of the group, and we spell out the implications for improved data collection in §5.

[5] Following standard convention, we assign Hispanic as the target variable value as long as a person self-identified as Hispanic for their ethnicity, regardless of their self-reported race.

[6] Alternatively, in some instances (*i.e.*, anonymized datasets), what we define as baseline features may be the least likely to be available. We replicate the main results without baseline features in Appendix F.

[7] We note that EHR records commonly drop accents from names, representing "José López" as "Jose Lopez."

[8] For each vital sign (*e.g.*, blood pressure), we collect all records associated with the patient in between 2010 and 2022. The weight associated with each measurement is defined as $w_i = \frac{T-t_i}{T}$ where w_i is the weight of the i-th value, t_i is the time from the measurement to the most recent measurement, and T is the time from the patient's earliest recorded measurement to the most recent measurement. The weights are adjusted such that $\Sigma w_i = 1$.

Table 1: Label and feature characteristics. For continuous features, the mean and variance are shown. For categorical features, the top-ranked modal categories and the corresponding proportion of subjects are shown.

		Mean (SD) / Proportion
Label	Race/ethnicity	White: 75.47%, Hispanic or Latino: 12.59%, Black or African American: 9.00%, Asian: 2.23%, American Indian and Alaska Native: 0.47%, Native Hawaiian and Other Pacific Islander: 0.23%
Features	**Baseline (k=21)**	
	First name	James: 0.013%, Michael: 0.012%, John: 0.011%, Robert: 0.011%, David: 0.010%, Mary: 0.010%, William: 0.009%, · · ·
	Surname	Smith: 0.010%, Johnson: 0.007%, Williams: 0.006%, Jones: 0.006%, Brown: 0.006%, Davis: 0.005%, Miller: 0.005%, · · ·
	Geography (census block group)	605301110510: 4.3E-04%, 310459507005: 3.7E-04%, 605301110320: 3.6E-04%, 605301110520: 3.2E-04%, · · ·
	Demographic (k=62)	
	Age	49.95 (23.51)
	Gender	Female: 55.26%, Male: 44.69%, Other: 0.05%
	Marital status	Married: 37.80%, Single: 30.86%, Other: 23.03%, Divorced/Separated: 4.84%, Widowed: 3.48%
	State	TX: 12.77%, CA: 7.20%, AR: 6.59%, VA: 5.11%, FL: 4.59%, IL: 3.90%, CO: 3.83%, NC: 3.54%, AL: 3.48%, OH: 2.40%, · · ·
	Behavioral (k=12)	
	Language preference	English: 8.97%, Multiple: 3.86%, Other: 1.66%, Spanish: 0.18%, Vietnamese: 0.17%, Yiddish: 0.052%, Chinese: 0.035%, · · ·
	Tobacco	Has ever used tobacco or smoked: 30.30%
	Administrative (k=32)	
	Insurance group	Had any: 89.89%, Had Blue Cross Blue Shield: 32.45%, Had Medicare: 21.73%, Had UnitedHealthcare: 14.06%, · · ·
	Visit count per year	2010: 0.0387 (0.5003), 2011: 0.0645 (0.6035), 2012: 0.1354 (0.9317), 2013: 0.1755 (1.0543), 2014: 0.3839 (1.5000), · · ·
	General Health (k=48)	
	Allergies	No known drug allergy: 6.67%, Penicillin: 5.15%, No known allergy: 4.78%, Sulfa: 3.15%, Codeine: 2.07%, · · ·
	Immunizations count	0.29 (1.75)
	Vital signs	Body mass index: 28.39 (7.74), BMI observations: 12.23 (14.99), Systolic blood pressure: 124.03 (15.19), Systolic BP observations: 12.66 (16.52), · · ·
	Medical (k=1057)	
	Diagnosis codes	Z0000 (General adult examination): 42.86%, Z23 (Immunization): 38.42%, I10 (Hypertension): 34.28%, · · ·
	Procedure codes	Medicine services: 97.54%, Evaluation and management: 96.64%, Pathology: 63.61%, Surgery: 47.25%, · · ·

Table 2: Example $P(F|R)$, $P(R|S)$, and $P(G|R)$ probabilities. These are the probabilistic inputs to conventional BIFSG and the baseline features for the machine learning models.

Name or Census Block Group	*Probabilities*					
Michael	P(F|AIAN)=0.015	P(F|API)=0.009	P(F|BLACK)=0.015	P(F|HISP)=0.006	P(F|WHITE)=0.027	P(F|OTHER)=0.020
Smith	P(AIAN|S)=0.009	P(API|S)=0.005	P(BLACK|S)=0.231	P(HISP|S)=0.024	P(WHITE|S)=0.709	P(OTHER|S)=0.022
310459507005	P(G|AIAN)=8.9E-05	P(G|API)=5.6E-11	P(G|BLACK)=1.3E-07	P(G|HISP)=4.6E-07	P(G|WHITE)=5.1E-06	P(G|OTHER)=5.0E-06
Jose	P(F|AIAN)=0.002	P(F|API)=0.002	P(F|BLACK)=0.000	P(F|HISP)=0.047	P(F|WHITE)=0.000	P(F|OTHER)=0.001
Lopez	P(AIAN|S)=0.004	P(API|S)=0.010	P(BLACK|S)=0.006	P(HISP|S)=0.929	P(WHITE|S)=0.049	P(OTHER|S)=0.003
605301050130	P(G|AIAN)=4.7E-10	P(G|API)=5.6E-11	P(G|BLACK)=2.5E-11	P(G|HISP)=6.0E-06	P(G|WHITE)=3.5E-06	P(G|OTHER)=4.6E-06

codes are one-hot encoded. In total, this preprocessing yields 1,000 diagnosis codes and 55 procedure codes.

We also create a missing indicator for each feature, with a value of 1 if a subject has no records corresponding to that feature in the AFC dataset, and 0 otherwise. These missingness indicators capture social patterns that arise from lack of access to or utilization of health care services [7, 68, 96], underdiagnosis of medical conditions [21, 33, 34, 62, 89], lower rates of immunization [16, 43], and other racial, socioeconomic, or geographic causes for missingness.

We derive 1,232 features in total across all the feature categories, including missingness indicators. Additional details on feature cleaning and coding can be found in Appendix C.

3.3 Imputation and Training Procedure

We implement conventional BIFSG using the baseline features: first name priors, surname priors, and geography priors. The BIFSG posterior probability that a person belongs to a race R (including Other), given their first name F, surname S, and geographic area G, is given by the following equation where n is the number of racial categories:

$$P(R|S,F,G) = \frac{P(R|S) \cdot P(F|R) \cdot P(G|R)}{\sum_{R=1}^{n} P(R|S) \cdot P(F|R) \cdot P(G|R)}. \quad (1)$$

The study population is randomly divided into training and test sets with an 80%/20% split, stratified on the race label to preserve the proportion of each race group across each split. To impute race from AFC features, we train a random forest multi-class classifier. Optimal hyperparameters are established using a stratified five-fold cross validation on the training set. The model is trained using scikit-learn 1.0.2. For hyperparameter tuning detail, see Appendix E.

We train a random forest model with only baseline features, the same probabilistic inputs that are used by the conventional BIFSG model. We refer to this model as "ML BIFSG." This allows us to compare conventional BIFSG to an ML-based approach that uses the same underlying information along with observed race in the training data. For comparison with conventional BIFSG's racial categories, we use a single API category. We then train five additional models, incrementally adding demographic, behavioral, administrative, general health, and medical features as categorized in Table 1 in §3.2. We compare the performance metrics across the six models to understand how additional features impact the ability to impute race. For all models, we calculate 95% confidence intervals using 100 bootstrap replications on the test set predictions.

To assess performance, we calculate Area Under the Precision-Recall Curve (AUPRC), AUROC, F1-score, precision, and recall. We

focus on AUPRC due to large differences in the size of racial groups, as AUROC can present deceptively good performance on such imbalanced datasets [83]. The baseline of AUPRC is the fraction of positive cases, which is the prevalence of each race shown in Table 1. We present metrics at both the micro level – which weighs all individuals equally – and the macro level – which weighs racial subgroups equally. The latter is relevant for assessing disparities between subgroups.

4 RESULTS

We now provide results on (a) how ML performs relative to conventional BIFSG, (b) the impact of including more extensive features to test the redundant encoding hypothesis, (c) the ability to disaggregate the API category, and (d) the relative performance gains to increasing the feature set across subgroups.

4.1 Conventional and Machine Learning BIFSG Comparison

Table 3 and Figure 1 summarize performance metrics on the 20% test set, comparing conventional BIFSG and ML with the same inputs. Evaluation metrics by race for each model are in Appendix F.

First, we note that the conventional BIFSG model performs reasonably well, comparable to or better than BIFSG performance typically observed in other domains. BIFSG achieves a macro AUROC of 0.861 (95% CI, 0.859-0.863), which is much higher than the results from [64], which performs conventional BISG and BIFSG using voter data from five states and achieves an AUROC of 0.74 on an Alabama validation set. It is also slightly higher than the ML predictor that has an AUROC of 0.857 in [64]. It is comparable to the AUROC of 0.833 in [55] which uses a multilayer perceptron model with medical features derived from over 1.5 million unique patients' anonymized EHR. The fact that the conventional BIFSG model performs at least as well as other models that utilize many other features shows that name and geographic location information are quite informative. Figure 2, for instance, illustrates that using baseline features alone, shown with the "x" marker, outperforms using medical features alone.

Second, Table 3 shows that ML BIFSG improves performance over conventional BIFSG for every metric at both micro and macro levels. The difference between macro and micro performance reflects performance differences across very differently sized subgroups because macro performance weighs the performance for minority groups equally to performance for the majority White group, while micro performance is dominated by the performance on the White population. ML BIFSG improves upon conventional BIFSG by the largest margin at the macro level, which is most relevant to disparity assessments. The ML approach has a macro AUPRC at 0.666 (95% CI, 0.663-0.668) compared to 0.597 (95% CI, 0.595-0.599) with the conventional approach.

Third, to further understand what is driving the improvement, we investigate the performance metrics by race and find that the ML approach improves upon conventional BIFSG across all racial groups, with the most significant improvement for smaller racial minorities. Figure 1 compares conventional BIFSG with ML BIFSG, with the former on the x-axis and the latter on the y-axis. Each dot represents the corresponding metrics for each race group, and the 45

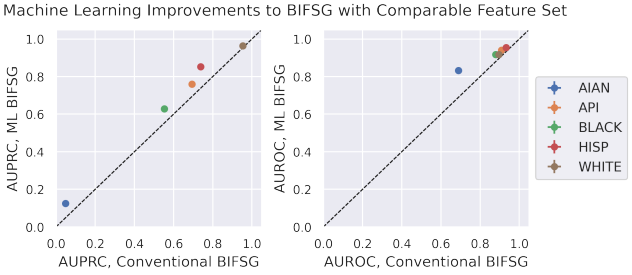


Figure 1: Machine learning improvements to BIFSG with a comparable feature set by race, with AUPRC on the left and AUROC on the right. Error bars representing 95% confidence intervals are too tight to be seen. Switching from conventional BIFSG to ML-based imputation using BIFSG probability priors yields better performance on both AUPRC and AUROC (calculated as one vs. rest) for all racial groups. Performance improves for all racial groups with the largest gain for minority groups.

degree line would indicate identical performance. All race groups lie on or above the 45 degree line, indicating that they all improve upon AUPRC and AUROC by switching from conventional to ML BIFSG. According to Table 6 in Appendix F, when comparing the difference in means, White patients have the smallest increase in AUPRC, from 0.953 (95% CI, 0.953-0.954) to 0.964 (95% CI, 0.964-0.964). Hispanic patients have the greatest increase in AUPRC, from 0.738 (95% CI, 0.735-0.741) to 0.853 (95% CI, 0.851-0.855), followed by AIAN and Black patients. In terms of AUROC, we see slight improvements in AUROC for White, Hispanic, Black, and API groups. AIAN subjects see the greatest increase in AUROC from 0.690 (95% CI, 0.682-0.696) with conventional BIFSG to 0.832 (95% CI, 0.825-0.837) with ML BIFSG. Considering that the conventional BI(F)SG technique has exhibited the poorest performance for AIAN and Black subjects [2, 24, 95], these performance gains are important [22].

Last, we conduct several analyses in Appendix G to better understand the mechanism through which ML improves upon conventional BIFSG. We show that ML does not simply improve upon conventional BIFSG by adjusting for the difference in base rates between the Census and the AFC patient population. We also find that the boost does not seem to stem from simple interaction terms and relaxing the (Naive Bayes) independence assumption. Instead, we find the most support for performance improvement from modeling non-linearities in the feature set, as displayed in Figure 5.

4.2 Machine Learning Performance with Increasing Feature Set

We now assess the impact of features beyond BIFSG inputs. Figure 2 displays the micro and macro AUPRC and AUROC for six random forest models to incrementally expand the feature set. The x-axis shows the feature set on which each of the six models was trained, beginning with baseline features alone, and incrementally adding categories of features to the prior feature set. The performance from a model that uses all previous feature sets and from a model that uses only the current feature set are shown by the "o" and "x"

Table 3: Micro and macro machine learning improvements to BIFSG with comparable feature set, with 95% confidence intervals.

Model	*AUPRC*	*AUROC*	*F1-Score*	*Precision*	*Recall*
			Micro		
Conventional BIFSG	0.898 (0.898-0.899)	0.967 (0.967-0.967)	0.864 (0.864-0.865)	0.864 (0.864-0.865)	0.864 (0.864-0.865)
ML BIFSG	0.934 (0.933-0.934)	0.980 (0.980-0.980)	0.881 (0.881-0.881)	0.881 (0.881-0.881)	0.881 (0.881-0.881)
			Macro		
Conventional BIFSG	0.597 (0.595-0.599)	0.861 (0.859-0.863)	0.596 (0.594-0.599)	0.677 (0.673-0.683)	0.560 (0.558-0.562)
ML BIFSG	0.666 (0.663-0.668)	0.912 (0.911-0.913)	0.620 (0.617-0.621)	0.813 (0.802-0.825)	0.581 (0.580-0.583)

markers respectively; Table 7 in Appendix F reports the numerical values of these metrics. We note that all metrics have very tight confidence intervals.

As we add in more feature sets, we observe a steady but very small increase in both micro and macro AUPRC and AUROC as the blue and orange lines trend slightly upwards but stay relatively flat in Figure 2. There is only a marginal improvement of micro AUPRC from 0.933 (95% CI, 0.933-0.934) with baseline features alone to 0.944 (95% CI, 0.943-0.944) with the full feature set. The addition of demographic features produces the largest increase in performance from 0.933 to 0.939 (95% CI, 0.938-0.939). For macro AUPRC, we observe an increase with more features from 0.585 (95% CI, 0.581-0.588) to 0.621 (95% CI, 0.618-0.625). Turning to AUROC, we see nearly no change in micro AUROC, from 0.983 (95% CI, 0.983-0.983) with baseline features alone to 0.985 (95% CI, 0.985-0.986) with the full feature set, and a small increase in macro AUROC from 0.893 (95% CI, 0.891-0.896) to 0.907 (95% CI, 0.906-0.909). Moreover, we observe that baseline features alone yield the best performance followed by demographics, medical, general health, administrative, and behavioral data, tempering the concern that unexpected data sources will recover race as well as features that are most closely related to race such as name, geography, and demographics. We provide a similar analysis in Appendix F in a setting where baseline features are not available (*i.e.*, anonymized datasets) and find that models trained with all other feature sets never reach the same performance as using name and geography probabilistic priors. This shows that while including additional features does encode further probabilistic information about race, using EHR-derived information excluding name and geography probabilistic priors still results in a significant amount of uncertainty.

The small performance boost is surprising from the perspective of redundant encodings. Adding over 1,000 rich features, including more demographic information, insurance groups, health care visits, medical diagnoses, and procedures, yields no substantive prediction gain in addition to the baseline features, and nothing close to the textbook exposition that posits that the probability of correctly inferring race should approach one. This finding is highly relevant for understanding redundant encodings in the wild. First, redundant encodings are less about "big data" than the underlying quality of the data. In this setting, name and geography appear to be the most highly relevant features, affirming the value of BIFSG against existing criticisms. Second, we observe only modest evidence that race imputations improve as EHR information is added, notwithstanding extensive research documenting disparities in health conditions that are encoded in our medical diagnostic codes. In contrast to textbook theoretical examples, our evidence

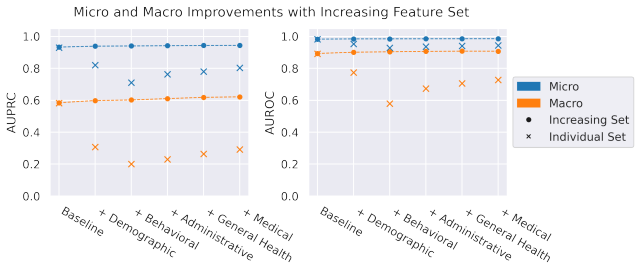


Figure 2: Micro and macro performance for an increasing feature set, with AUPRC on the left and AUROC on the right. "x" markers indicate the micro and macro performance for each feature set alone. Error bars representing 95% confidence intervals are too tight to be seen. There is a steady but very small increase in both micro and macro AUPRC and AUROC when incorporating additional features (increasing set). Baseline features alone yield the best performance followed by demographics, medical, general health, administrative, and behavioral data (individual set).

suggests that each additional feature is not an *independent* signal of race. Third, this evidence does not corroborate the sharpest claims that big data violate anti-discrimination law's anti-classification principle, as race remains measured with substantial uncertainty, particularly for smaller racial groups.

4.3 Disaggregating the API Category

While many administrative datasets follow the 1977 OMB standard and use the API category, we have access to the underlying Asian and NHPI labels which cannot be used within conventional BIFSG. We now disaggregate our predictions to present results of six race and ethnicity categories instead of five as in §4.2, which simulates the impact of OMB's revision of race reporting standards in 1997. Here we focus on the impact of the disaggregating the API category, before turning to insights across all six subgroups in §4.4.

Table 4 and Table 8 in Appendix F show how precision and recall change as the feature sets increase, and illustrate that the subgroups of the API category experience quite different impacts of incorporating additional features. First, using only baseline features yields a precision of 0.829 (95% CI, 0.824-0.834) for Asian subjects and 0.645 (95% CI, 0.582-0.692) for NHPI subjects, and a recall of 0.722 (95% CI, 0.716-0.727) for Asian subjects and only 0.070 (95% CI, 0.059-0.080) for NHPI subjects. We then find that while the addition of general health data increases precision by only 0.4% (95% CI, 0.2%-0.6%) for Asian subjects, it increases precision by 7.4% (95%

CI, 2.4%-11.5%) for NHPI subjects. Similarly, while the addition of medical features changes recall slightly by -0.1% (95% CI, -0.2%-0.1%) for Asian subjects, it increases recall by 5.9% (95% CI, 1.2%-11.8%) for NHPI subjects. This exercise shows that further disaggregating race categories, as increasingly advocated by the medical community [52, 81, 100], in conjunction with ML-based approaches can unlock more actionable and valuable insights. As illustrated in Appendix H, such disaggregation matters. Conventional BIFSG estimates the prevalence of asthma diagnostics, for instance, as 620 (per 10,000) API patients. But this masks dramatic differences, with the NHPI rate of 1,363 nearly twice that for Asians. Although estimates are noisy, ML-based approaches are able to substantially disaggregate and detect those dramatic disparities within the API category.

4.4 Performance by Racial Groups with Increasing Feature Set

We now investigate the performance gains across all subgroups from increasing the feature set. We demonstrate a more nuanced effect where the trade-offs between precision and recall can differ in subtle ways for different groups under prevailing practice.

Figure 3 illustrates the changes in and trade-offs between precision and recall measures with each incremental feature category added at the 0.5 threshold for each racial group. The color of each marker signifies the race of each group. The size of each marker is proportional to the prevalence of the racial group. The transparency of the circle represents the richness of the features used to generate the performance (*e.g.*, the most transparent circle represents the baseline features, while the non-transparent circle represents the full feature set). For instance, for the green dots representing the prediction performance of Black patients, following the trace from the most to least transparent performance markers going from bottom left to top right, we observe a steady increase in precision and recall as we add in demographics, behavioral, administrative, general health, and medical features. According to Table 8 in Appendix F, which shows the values of precision and recall, we observe a cumulative 10% change in recall from 0.401 (95% CI, 0.398-0.405) to 0.442 (95% CI, 0.439-0.445) as we add in demographic, behavioral, administrative, general health, and medical features, as well as a cumulative 6% change in precision from 0.742 (95% CI, 0.740-0.746) to 0.788 (95% CI, 0.784-0.791) with the addition of demographic to medical features.

We find that the impacts on precision and recall are different for racial groups as more features are incorporated. White patients experience the smallest cumulative changes with less than 1% change in precision and less than 2% change in recall as we increase the feature set from baseline to medical. Asian subjects also see less than 2% cumulative changes in precision and recall, while, as we previously highlighted, NHPI subjects disaggregated from the same shared API category see larger cumulative changes of 3% in precision and 6% in recall. Though Hispanic subjects see very little change in recall, there is a more substantial cumulative change of 6% increase in precision. For Black and AIAN subjects, we see increases in both precision and recall, with demographic features providing the greatest increase for the Black group, and administrative and general health features providing the greatest increases for the AIAN group.

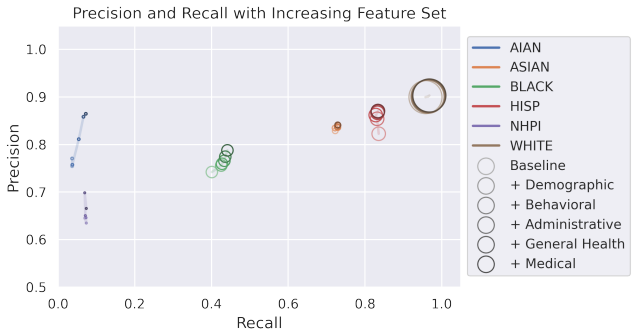


Figure 3: Precision and recall at the 0.5 threshold for an increasing feature set for each racial group. The color of each marker signifies the race of each group, and the size of each marker is proportional to the prevalence of the racial group. Following the gradient of the circles allows us to track the precision and recall trade-offs for each racial group. For example, while we see an increase in precision and recall when adding features for Black patients, we observe much smaller magnitudes of change in the same metrics for White patients.

These findings add important considerations regarding data quality and trade-offs between performance metrics when we use more data to conduct racial imputation. While we have shown in §4.2 that additional features do not substantially improve imputation overall, they can impact subgroups to substantially different degrees, particularly smaller racial minorities. This finding contributes to the empirical evidence that increased dataset size does not necessarily close all performance gaps between subgroups, echoing [25]. That being said, we find that additional features do sharpen our ability to discern differences between some subgroups, as we demonstrated with the dramatic health disparities detected between Asian and NHPI – a demonstration, also, of the importance of disaggregation of race groups into more granular subgroups.

5 DISCUSSION

In this paper, we have presented evidence that ML can improve inferences about race using name and census geography information, but that additional and richer demographic and health care features from a highly realistic EHR dataset yield only marginal improvements. These findings address one of the most widespread conjectures in algorithmic fairness, namely that redundant encodings with ML and big data threaten a core tenet of anti-discrimination law. Our results highlight the encoding of race information in name and geography, and provide important context on redundant encodings in the health care setting. Redundant encodings are less a function of the *size* of data, than of specific informative inputs. These findings also reveal the differential gains that a large feature set can have for minority groups, with different precision and recall trade-offs for each racial groups, as well as for each category of data with varying levels of quality. We have also shown how ML can enable data disaggregation of the API group, enabling researchers to identify important health disparities between Asian and NHPI subgroups [37, 46, 100].

Table 4: Percent change in performance for each increasing feature set by race, with 95% confidence intervals.

(a) Percent change in precision.

Features	*AIAN*	*ASIAN*	*BLACK*	*HISP*	*NHPI*	*WHITE*
Baseline	–	–	–	–	–	–
+ Demographic	2.2% (0.6%-5.0%)	0.7% (0.5%-0.9%)	1.8% (1.5%-2.2%)	3.8% (3.7%-4.0%)	-1.6% (-8.1%-4.1%)	0.1% (0.1%-0.2%)
+ Behavioral	-1.7% (-4.0%-0.5%)	0.1% (-0.0%-0.2%)	0.5% (0.3%-0.7%)	0.9% (0.8%-1.0%)	1.8% (-2.1%-7.7%)	-0.0% (-0.0%- -0.0%)
+ Administrative	7.1% (3.7%-11.8%)	0.2% (0.0%-0.3%)	0.8% (0.6%-1.1%)	0.6% (0.5%-0.6%)	0.7% (-3.3%-4.9%)	0.2% (0.1%-0.2%)
+ General Health	5.8% (3.8%-8.0%)	0.4% (0.2%-0.6%)	1.1% (0.9%-1.4%)	0.3% (0.2%-0.4%)	7.4% (2.4%-11.5%)	0.0% (0.0%-0.1%)
+ Medical	0.7% (-0.7%-2.2%)	0.1% (-0.0%-0.2%)	1.7% (1.5%-1.9%)	0.1% (0.1%-0.2%)	-4.7% (-8.7%- -1.8%)	0.1% (0.1%-0.1%)

(b) Percent change in recall.

Features	*AIAN*	*ASIAN*	*BLACK*	*HISP*	*NHPI*	*WHITE*
Baseline	–	–	–	–	–	–
+ Demographic	0.0% (-2.2%-3.2%)	-0.1% (-0.3%-0.2%)	6.1% (5.7%-6.6%)	-0.5% (-0.6%- -0.4%)	5.9% (1.1%-12.1%)	0.6% (0.5%-0.6%)
+ Behavioral	1.7% (-1.8%-5.4%)	1.0% (0.8%-1.2%)	0.5% (0.2%-0.7%)	-0.4% (-0.5%- -0.4%)	-1.1% (-5.7%-4.3%)	0.2% (0.1%-0.2%)
+ Administrative	44.8% (34.7%-58.2%)	-0.2% (-0.3%- -0.0%)	1.4% (1.2%-1.7%)	0.6% (0.5%-0.6%)	-2.3% (-7.6%-3.6%)	0.1% (0.1%-0.1%)
+ General Health	22.5% (17.3%-30.6%)	0.2% (0.0%-0.4%)	0.5% (0.2%-0.9%)	0.1% (0.0%-0.1%)	-2.3% (-8.4%-3.8%)	0.1% (0.1%-0.1%)
+ Medical	9.3% (4.9%-16.7%)	-0.1% (-0.2%-0.1%)	1.3% (1.0%-1.6%)	0.1% (0.0%-0.1%)	5.9% (1.2%-11.8%)	0.1% (0.1%-0.1%)

We note several limitations to our work. First, our results do not allow us to reject the presence of redundant encodings in other settings. That said, the AFC dataset is a rich, real-world, and highly relevant dataset for health care, representing the actual information that health care providers utilize to algorithmically improve care [69, 87, 101]. Second, because EHR information is complex, and ours is the first to structure the AFC dataset for this kind of analysis, we are unable to use all available information. This is because EHR data are primarily used for efficient record-keeping for health care systems, and EHR data in informatics and ML research applications often require extensive pre-processing and cleaning (see Appendix C). That said, our analysis focuses on highly prevalent and important features that provide a comprehensive summary of subjects' demographics, patient history, and interactions with the health care system. Third, our analysis has only embarked on a demonstration by disaggregating the Asian and NHPI subgroups in the API category. The aggregation of widely varying subgroups into the broad race groups defined by OMB is a pervasive issue that extends far beyond the API group or the initial disaggregation we have offered, affecting many other subgroups across all races and ethnicities [6, 37, 46, 86]. As we show in Appendix B, our disaggregation of the API group is based on the OMB's 1997 revision [72], but even this reform was limited, given that the underlying EHR data contains richer information about subgroups for some health care providers. One of the main limiting factors is that Census tables that present the racial demographics of names and geographic regions, for example, still use the coarse 1977 OMB standard. Future work on disaggregation for racial inference and measurement, therefore, depends heavily on the U.S. Census and OMB to collect and report subgroup data.

Despite these limitations, our results provide evidence that theoretical concerns about redundant encodings – which are a mainstay in the algorithmic fairness literature – may not hold in practice across all settings in the way conventionally posited, including our rich data setting of health care delivery. At minimum, our findings suggest that concerns about redundant encodings be spelled out with greater specificity. We find that ML applied to a rich feature set does not capture race information substantially *beyond what is already encoded in name and geography characteristics.* As a corollary, the importance of name and geography features highlights that the conventional BI(F)SG methods do in fact already encapsulate high-quality information for race inference, which supports the use of these approaches to measure racial disparities. We provide evidence that a ML approach can improve inferences using the same feature set as for BI(F)SG methods. This finding dispels some skepticism of conventional BI(F)SG. When the Consumer Financial Protection Bureau used such methods to identify discrimination, for instance, critics charged these methods as "junk science" [56]. In practice, such imputation methods are increasingly relevant to measure racial disparities, as mandated by law, when race cannot be observed [29].

Even though the incorporation of richer features provides limited improvements over conventional BI(F)SG with moderated risk of redundant encoding, this does not mean that rich feature sets can be used indiscriminately. As illustrated in Appendix F, additional features used in increasing number do encode further probabilistic information about race, approaching the performance obtained from using name and geography features alone, albeit with significant uncertainty. Additionally, the varying impacts on precision and recall for the least-represented minority groups compared to the majority White group as the feature set grows is a subtle, yet important, effect. This impact can be hidden by cruder overall metrics like AUROC and micro-based metrics, so we emphasize a need for careful monitoring of more granular, group-specific metrics to identify such impacts of large feature sets. Of course, these effects can be best understood when racial groups are further disaggregated; for example, we observe significant differences in precision and recall between the disaggregated Asian and NHPI groups of the API category. Our ability to disaggregate the API group at all – which conventional BI(F)SG cannot do with the current name and geography tables – highlights the ability of ML to advance data disaggregation. Fundamentally, though, this points to the importance of improving race collection and reporting standards. A key area of future work is in developing name and geography tables

with disaggregated categories; such efforts can leverage the AFC dataset and other auxiliary datasets with more granular subgroup information to enable further our understanding of demographic disparities.

In sum, our work illustrates how ML with a rich feature set can improve the measurement of racial disparities, demonstrates these concerns in a rich, new dataset of health care practice, and shows that a widely cited theoretical concern of redundant encodings as undercutting anti-discrimination principles may not in practice operate across all settings, at least as conventionally posited as being about data size.

ACKNOWLEDGMENTS

We are grateful to the Stanford Institute for Human-Centered Artificial Intelligence and the Public Interest Technology University Network for supporting this research. We acknowledge the American Board of Family Medicine and the ABFM PRIME Registry participating clinicians, without whom the American Family Cohort would not be possible. We thank Isabella Chu, David Rehkopf, Bob Phillips, Ayin Vala, Esther Velasquez, and Shiying Hao for their support with accessing and understanding the American Family Cohort data; Cameron Raymond for research assistance; and Cam Guage, Benji Lu, and Jia Wan for helpful feedback.

REFERENCES

[1] Hammaad Adam, Ming Ying Yang, Kenrick Cato, Ioana Baldini, Charles Senteio, Leo Anthony Celi, Jiaming Zeng, Moninder Singh, and Marzyeh Ghassemi. 2022. Write It Like You See It: Detectable Differences in Clinical Notes by Race Lead to Differential Model Recommendations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 7–21. https://doi.org/10.1145/3514094.3534203

[2] Dzifa Adjaye-Gbewonyo, Robert A Bednarczyk, Robert L Davis, and Saad 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 (2014), 268–283.

[3] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing black-box models for indirect influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.

[4] Larry Alexander and Kevin Cole. 1997. Discrimination by Proxy. , 453–463 pages. https://scholarship.law.umn.edu/concomm/602

[5] American Hospital Association. 2022. ICD-10-CM Coding for Social Determinants of Health. https://www.aha.org/system/files/2018-04/value-initiative-icd-10-code-social-determinants-of-health.pdf

[6] Adrian Matias Bacong, Christina Holub, and Liki Porotesano. 2016. Comparing obesity-related health disparities among Native Hawaiians/Pacific Islanders, Asians, and whites in California: reinforcing the need for data disaggregation and operationalization. *Hawai'i Journal of Medicine & Public Health* 75, 11 (2016), 337.

[7] Zinzi D Bailey, Nancy Krieger, Madina Agénor, Jasmine Graves, Natalia Linos, and Mary T Bassett. 2017. Structural racism and health inequities in the USA: evidence and interventions. *The Lancet* 389, 10077 (2017), 1453–1463. https://doi.org/10.1016/S0140-6736(17)30569-X

[8] Asha Banerjee. 2022. Understanding economic disparities within the AAPI Community. https://www.epi.org/blog/understanding-economic-disparities-within-the-aapi-community/

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[10] Ann Barta, Gale McNeill, Peggy Meli, Kathleen Wall, and Ann Zeisset. 2008. ICD-10-CM primer. https://library.ahima.org/doc?oid=106177#.Y7YIzHbMLIU

[11] Ruha Benjamin. 2019. *Race After Technology: Abolitionist Tools for the New Jim Code*. Polity, Cambridge, UK.

[12] Claire Bowen, Aaron R Williams, and Ajjit Narayanan. 2021. To advance racial equity, releasing disaggregated data while Protecting Privacy will be key. https://www.urban.org/urban-wire/advance-racial-equity-releasing-disaggregated-data-while-protecting-privacy-will-be-key

[13] Consumer Financial Protection Bureau. 2018. 12 C.F.R. §1002.5(b).

[14] United States Census Bureau. 2022. American Community Survey Data via API. https://www.census.gov/programs-surveys/acs/data/data-via-api.html

[15] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. 2019. Fairness Under Unawareness: Assessing Disparity When Protected Class Is Unobserved. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 339–348. https://doi.org/10.1145/3287560.3287594

[16] Judy Y Chen, Sarah A Fox, Clairessa H Cantrell, Susan E Stockdale, and Marjorie Kagawa-Singer. 2007. Health disparities and prevention: racial/ethnic barriers to flu vaccinations. *Journal of community health* 32, 1 (2007), 5–20.

[17] U.S. Code. 1974. 15 U.S.C. §1691(a).

[18] Consumer Financial Protection Bureau. 2014. Using publicly available information to proxy for unidentified race and ethnicity. https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf

[19] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. https://doi.org/10.48550/ARXIV.1808.00023

[20] Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. 2017. Proxy Non-Discrimination in Data-Driven Systems. https://doi.org/10.48550/ARXIV.1707.08120

[21] Carla M. Davis, Andrea J. Apter, Adrian Casillas, Michael B. Foggs, Margee Louisias, Elsie C. Morris, Anil Nanda, Michael R. Nelson, Princess U. Ogbogu, Cheryl Lynn Walker-McGill, Julie Wang, and Tamara T. Perry. 2021. Health disparities in allergic and immunologic conditions in racial and ethnic underserved populations: A Work Group Report of the AAAAI Committee on the Underserved. *Journal of Allergy and Clinical Immunology* 147, 5 (2021), 1579–1593. https://doi.org/10.1016/j.jaci.2021.02.034

[22] Ari Decter-Frain. 2022. How should we proxy for race/ethnicity? Comparing Bayesian improved surname geocoding to machine learning methods. https://doi.org/10.48550/ARXIV.2206.14583

[23] Kevin DeLuca and John A. Curiel. 2022. Validating the Applicability of Bayesian Inference with Surname and Geocoding to Congressional Redistricting. *Political Analysis* (2022), 1–7. https://doi.org/10.1017/pan.2022.14

[24] Stephen F. Derose, Richard Contreras, Karen J. Coleman, Corinna Koebnick, and Steven J. Jacobsen. 2012. Race and ethnicity data quality and imputation using U.S. Census data in an integrated health system. *Medical Care Research and Review* 70, 3 (2012), 330–345. https://doi.org/10.1177/1077558712466293

[25] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. 2021. Retiring Adult: New Datasets for Fair Machine Learning. https://doi.org/10.48550/ARXIV.2108.04884

[26] Grant Duffy, Shoa L Clarke, Matthew Christensen, Bryan He, Neal Yuan, Susan Cheng, and David Ouyang. 2022. Confounders mediate AI prediction of demographics in medical imaging. *npj Digital Medicine* 5, 1 (2022), 1–6.

[27] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[28] Marc N Elliott, Peter A Morrison, Allen Fremont, Daniel F McCaffrey, Philip Pantoja, and Nicole 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 (2009), 69–83.

[29] Exec. Order. 13985. 2021. Exec order no. 13985 86 fed. reg. 7009, advancing racial equity and support for underserved communities through the federal government.

[30] Facebook. 2023. An Update on Our Ads Fairness Efforts. Facebook Newsroom. https://about.fb.com/news/2023/01/an-update-on-our-ads-fairness-efforts/

[31] Center for Behavioral Health Statistics and Quality. 2021. Racial/Ethnic Differences in Substance Use, Substance Use Disorders, and Substance Use Treatment Utilization among People Aged 12 or Older (2015-2019). *Publication No. PEP21-07-01-001* Rockville, MD: Substance Abuse and Mental Health Services Administration (2021), 12–87. https://www.samhsa.gov/data/sites/default/files/reports/rpt35326/2021NSDUHSUChartbook102221B.pdf

[32] The Annie E. Casey Foundation. 2016. Disaggregating data to find racial inequities. https://www.aecf.org/blog/taking-data-apart-why-a-data-driven-approach-matters-to-race-equity

[33] H Jack Geiger. 2003. *Racial and ethnic disparities in diagnosis and treatment: a review of the evidence and a consideration of causes.* National Academies Press, Washington, DC. https://www.ncbi.nlm.nih.gov/books/NBK220337/

[34] Kan Z. Gianattasio, Christina Prather, M. Maria Glymour, Adam Ciarleglio, and Melinda C. Power. 2019. Racial disparities and temporal trends in dementia misdiagnosis risk in the United States. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* 5 (2019), 891–898. https://doi.org/10.1016/j.trci.2019.11.008

[35] Judy Wawira Gichoya, Imon Banerjee, Ananth Reddy Bhimireddy, John L Burns, Leo Anthony Celi, Li-Ching Chen, Ramon Correa, Natalie Dullerud, Marzyeh Ghassemi, Shih-Cheng Huang, Po-Chih Kuo, Matthew P Lungren, Lyle J Palmer, Brandon J Price, Saptarshi Purkayastha, Ayis T Pyrros, Lauren Oakden-Rayner, Chima Okechukwu, Laleh Seyyed-Kalantari, Hari Trivedi, Ryan Wang, Zachary Zaiman, and Haoran Zhang. 2022. AI recognition of patient race in medical imaging: a modelling study. *The Lancet Digital Health* 4, 6 (June 2022), e406–e414. https://doi.org/10.1016/S2589-7500(22)00063-2

[36] Talia B Gillis. 2021. The input fallacy. *Minn. L. Rev.* 106 (2021), 1175.

[37] Nancy P Gordon, Teresa Y Lin, Jyoti Rau, and Joan C Lo. 2019. Aggregation of Asian-American subgroups masks meaningful differences in health and health risks among Asian ethnicities: an electronic health record based cohort study. *BMC public health* 19, 1 (2019), 1–14.

[38] GovInfo. 2022. Privacy act of 1974. https://www.govinfo.gov/content/pkg/USCODE-2018-title5/pdf/USCODE-2018-title5-partI-chap5-subchapII-sec552a.pdf

[39] Larry A. Green, George E. Fryer, Barbara P. Yawn, David Lanier, and Susan M. Dovey. 2001. The Ecology of Medical Care Revisited. *New England Journal of Medicine* 344, 26 (2001), 2021–2025. https://doi.org/10.1056/NEJM200106283442611 arXiv:https://doi.org/10.1056/NEJM200106283442611 PMID: 11430334.

[40] Robert W. Grundmeier, Lihai Song, Mark J. Ramos, Alexander G. Fiks, Marc N. Elliott, Allen Fremont, Wilson Pace, Richard C. Wasserman, and Russell Localio. 2015. Imputing missing race/ethnicity in pediatric electronic health records: Reducing bias with use of U.S. Census location and surname Data. *Health Services Research* 50, 4 (2015), 946–960. https://doi.org/10.1111/1475-6773.12295

[41] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 501–512. https://doi.org/10.1145/3351095.3372826

[42] Moritz Hardt. 2016. How big data is unfair. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

[43] Amelia M Haviland, Marc N Elliott, Katrin Hambarsoomian, and Nicole Lurie. 2011. Immunization disparities by Hispanic ethnicity and language preference. *Archives of internal medicine* 171, 2 (2011), 158–165.

[44] Daniel E Ho and Alice Xiang. 2020. Affirmative algorithms: The legal grounds for fairness as awareness. , 134 pages.

[45] Kelly M Hoffman, Sophie Trawalter, Jordan R Axt, and M Norman Oliver. 2016. Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites. *Proceedings of the National Academy of Sciences* 113, 16 (2016), 4296–4301.

[46] Ariel T Holland and Latha P Palaniappan. 2012. Problems with the collection and interpretation of Asian-American health data: omission, aggregation, and extrapolation. *Annals of epidemiology* 22, 6 (2012), 397–405.

[47] Kosuke Imai and Kabir Khanna. 2016. Improving ecological inference by predicting individual ethnicity from Voter Registration Records. *Political Analysis* 24, 2 (2016), 263–272. https://doi.org/10.1093/pan/mpw001

[48] Kosuke Imai, Santiago Olivella, and Evan T. R. Rosenman. 2022. Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. *Science Advances* 8, 49 (2022), eadc9824. https://doi.org/10.1126/sciadv.adc9824 arXiv:https://www.science.org/doi/pdf/10.1126/sciadv.adc9824

[49] Michael E. Johansen, Sheetal M. Kircher, and Timothy R. Huerta. 2016. Reexamining the Ecology of Medical Care. *New England Journal of Medicine* 374, 5 (2016), 495–496. https://doi.org/10.1056/NEJMc1506109 arXiv:https://doi.org/10.1056/NEJMc1506109 PMID: 26840150.

[50] Nathan Joo, Richard V. Reeves, and Edward Rodrigue. 2022. https://www.brookings.edu/research/asian-american-success-and-the-pitfalls-of-generalization/

[51] Nathan Kallus, Xiaojie Mao, and Angela Zhou. 2021. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science* 68, 3 (Apr 2021), 1959–1981. https://doi.org/10.1287/mnsc.2020.3850

[52] Tina J Kauh, Jen'nan Ghazal Read, and AJ Scheitler. 2021. The critical role of racial/ethnic data disaggregation for health equity. *Population research and policy review* 40, 1 (2021), 1–7.

[53] Kenneth G Keppel, Jeffrey N Pearcy, Diane K Wagener, et al. 2002. Trends in racial and ethnic-specific rates for the health status indicators: United States, 1990-98. http://www.cs.cmu.edu/~eugene/refs/f-trials/Keppel-al-02.pdf

[54] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding Discrimination through Causal Reasoning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 656–666.

[55] Ji-Sung Kim, Xin Gao, and Andrey Rzhetsky. 2018. RIDDLE: Race and ethnicity Imputation from Disease history with Deep LEarning. *PLOS Computational Biology* 14, 4 (April 2018), e1006106. https://doi.org/10.1371/journal.pcbi.1006106

[56] James Rufus Koren. 2016. Feds use Rand formula to spot discrimination. The GOP calls it junk science. https://www.latimes.com/business/la-fi-rand-elliott-20160824-snap-story.html.

[57] Katie Labgold, Sarah Hamid, Sarita Shah, Neel R Gandhi, Allison Chamberlain, Fazle Khan, Shamimul Khan, Sasha Smith, Steve Williams, Timothy L Lash, et al. 2021. Estimating the unknown: greater racial and ethnic disparities in COVID-19 burden after accounting for missing race/ethnicity data. *Epidemiology (Cambridge, Mass.)* 32, 2 (2021), 157.

[58] LaTasha Lee, Kim Smith-Whitley, Sonja Banks, and Gary Puckrein. 2019. Reducing health care disparities in sickle cell disease: a review. *Public Health Reports* 134, 6 (2019), 599–607.

[59] Jeffrey W Lockhart, Molly M King, and Christin Munsch. 2023. Name-based demographic inference and the unequal distribution of misrecognition. *Nature Human Behaviour* (2023), 1–12. https://doi.org/10.1038/s41562-023-01587-9

[60] Clara Lu, Rabeeyah Ahmed, Amel Lamri, and Sonia S Anand. 2022. Use of race, ethnicity, and ancestry data in health research. *PLOS Global Public Health* 2, 9 (2022), e0001060.

[61] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. 2021. On the Applicability of Machine Learning Fairness Notions. *SIGKDD Explor. Newsl.* 23, 1 (may 2021), 14–23. https://doi.org/10.1145/3468507.3468511

[62] A James Mamary, Jeffery I Stewart, Gregory L Kinney, John E Hokanson, Kartik Shenoy, Mark T Dransfield, Marilyn G Foreman, Gwendolyn B Vance, Gerard J Criner, COPDGene® Investigators, et al. 2018. Race and gender disparities are evident in COPD underdiagnoses across all severities of measured airflow obstruction. *Chronic Obstructive Pulmonary Diseases: Journal of the COPD Foundation* 5, 3 (2018), 177.

[63] J Mathew, C Hodge, and M Khau. 2020. Z codes utilization among Medicare Fee-for-Service (FFS) beneficiaries in 2017.

[64] Kasey Matthews, Piotr Zak, Austin Li, Christien Williams, Sean Kamkar, and Jay Budzik. 2022. Zest Race Predictor. https://github.com/zestai/zrp/blob/main/model_report.rst

[65] Vickie M. Mays, Susan D. Cochran, and Namdi W. Barnes. 2007. Race, race-based discrimination, and health outcomes among African Americans. *Annual Review of Psychology* 58, 1 (2007), 201–225. https://doi.org/10.1146/annurev.psych.57.

102904.190212
[66] National Bureau of Economic Research. 2022. ICD-9-CM to and from ICD-10-CM and ICD-10-PCS crosswalk or general equivalence mappings. https://www.nber.org/research/data/icd-9-cm-and-icd-10-cm-and-icd-10-pcs-crosswalk-or-general-equivalence-mappings
[67] National Library of Medicine. 2021. SNOMED CT to ICD-10-CM Map. https://www.nlm.nih.gov/research/umls/mapping_projects/snomedct_to_icd10cm.html
[68] Alan Nelson. 2002. Unequal treatment: confronting racial and ethnic disparities in health care. *Journal of the National Medical Association* 94, 8 (2002), 666.
[69] Kee Yuan Ngiam and Ing Wei Khor. 2019. Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology* 20, 5 (2019), e262–e273. https://doi.org/10.1016/S1470-2045(19)30149-4
[70] U.S. Department of Health and Human Services. 2021. Office of Minority Health Minority Population Profiles. https://www.minorityhealth.hhs.gov/
[71] U.S. Department of Health and Human Services. 2022. Data Brief: Inaccuracies in Medicare's Race and Ethnicity Data Hinder the Ability To Assess Health Disparities. https://oig.hhs.gov/oei/reports/OEI-02-21-00100.pdf
[72] Office of Management and Budget. 1997. Revisions to the standards for the classification of federal data on race and ethnicity. *Federal Register* 62, 210 (1997), 58782–58790.
[73] Sam S Oh, Joshua Galanter, Neeta Thakur, Maria Pino-Yanes, Nicolas E Barcelo, Marquitta J White, Danielle M de Bruin, Ruth M Greenblatt, Kirsten Bibbins-Domingo, Alan HB Wu, et al. 2015. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS medicine* 12, 12 (2015), e1001918.
[74] Karin Orvis. 2022. Reviewing and revising standards for maintaining, collecting, and presenting federal data on Race and ethnicity. https://www.whitehouse.gov/omb/briefing-room/2022/06/15/reviewing-and-revising-standards-for-maintaining-collecting-and-presenting-federal-data-on-race-and-ethnicity/
[75] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-Aware Data Mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Las Vegas, Nevada, USA) *(KDD '08)*. Association for Computing Machinery, New York, NY, USA, 560–568. https://doi.org/10.1145/1401890.1401959
[76] James Price, Jagdish Khubchandani, Molly Mckinney, and Robert Braun. 2013. Racial/Ethnic Disparities in Chronic Diseases of Youths and Access to Health Care in the United States. *BioMed Research International* 2013 (01 2013), 787616. https://doi.org/10.1155/2013/787616
[77] Anya Prince and Daniel Schwarcz. 2020. Proxy Discrimination in the Age of Artificial Intelligence and Big Data. https://ilr.law.uiowa.edu/print/volume-105-issue-3/proxy-discrimination-in-the-age-of-artificial-intelligence-and-big-data/
[78] Megan Randall, Alena Stern, and Yipeng Su. 2021. Five ethical risks to consider before filling missing race and ethnicity data. https://www.urban.org/research/publication/five-ethical-risks-consider-filling-missing-race-and-ethnicity-data
[79] Prime Registry. 2022. About PRIME Registry. https://primeregistry.org/
[80] Aaron Rieke, Vincent Southerland, Dan Svirsky, and Mingwei Hsu. 2022. Imperfect Inferences: A Practical Assessment. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 767–777. https://doi.org/10.1145/3531146.3533140
[81] Victor Rubin, Danielle Ngo, A Ross, Dalila Butler, and Nisha Balaram. 2018. Counting a diverse nation: Disaggregating data on race and ethnicity to advance a culture of health. https://www.policylink.org/resources-tools/counting-a-diverse-nation
[82] Georgia Robins Sadler, Lisa Ryujin, Tammy Nguyen, Gia Oh, Grace Paik, and Brenda Kustin. 2003. Heterogeneity within the Asian American community. *International Journal for Equity in Health* 2, 1 (2003), 1–9.
[83] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PloS one* 10, 3 (2015), e0118432.
[84] Reva B Siegel. 2003. Equality talk: Antisubordination and anticlassification values in constitutional struggles over Brown. *Harv. L. Rev.* 117 (2003), 1470.
[85] Michael W Sjoding, Robert P Dickson, Theodore J Iwashyna, Steven E Gay, and Thomas S Valley. 2020. Racial bias in pulse oximetry measurement. *New England Journal of Medicine* 383, 25 (2020), 2477–2478.
[86] Shobha Srinivasan and Tessie Guillermo. 2000. Toward improved health: disaggregating Asian American and Native Hawaiian/Pacific Islander data. *American journal of public health* 90, 11 (2000), 1731.
[87] Joshua D. Stein, Moshiur Rahman, Chris Andrews, Joshua R. Ehrlich, Shivani Kamat, Manjool Shah, Erin A. Boese, Maria A. Woodward, Jeff Cowall, Edward H. Trager, Prabha Narayanaswamy, and David A. Hanauer. 2019. Evaluation of an Algorithm for Identifying Ocular Conditions in Electronic Health Record Data. *JAMA Ophthalmology* 137, 5 (05 2019), 491–497. https://doi.org/10.1001/jamaophthalmol.2018.7051
[88] Thornburg v. Gingles. 1986. 478 U.S. 30.
[89] Patricia A Thomas. 2007. Racial and ethnic differences in osteoporosis. *JAAOS-Journal of the American Academy of Orthopaedic Surgeons* 15 (2007), S26–S30.
[90] Victoria Tran. 2018. https://www.urban.org/urban-wire/asian-americans-are-falling-through-cracks-data-representation-and-social-services
[91] Michael Carl Tschantz. 2022. What is Proxy Discrimination?. In *2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1993–2003. https://doi.org/10.1145/3531146.3533242
[92] Konstantinos (Office Of The Comptroller Of The Currency) Tzioumis. 2017. Data for: Demographic aspects of first names. https://doi.org/10.7910/DVN/TYJKEZ Type: dataset.
[93] US Census Bureau. 2021. Decennial Census Surname Files (2010, 2000). https://www.census.gov/data/developers/data-sets/surnames.html
[94] Ayin Vala, Shiying Hao, Isabella Chu, Robert LeRoy Phillips, and David Rehkopf. 2023. *The American Family Cohort (v12.2)*. Redivis, Stanford, CA. https://doi.org/10.57761/jn2e-7r28
[95] Ioan Voicu. 2018. Using First Name Information to Improve Race and Ethnicity Classification. *Statistics and Public Policy* 5, 1 (2018), 1–13. https://www.tandfonline.com/doi/full/10.1080/2330443X.2018.1427012
[96] Kellee White, Jennifer S Haas, and David R Williams. 2012. Elucidating the role of place in health care disparities: the example of racial/ethnic residential segregation. *Health services research* 47, 3pt2 (2012), 1278–1299.
[97] Kerr L. White, T. Franklin Williams, and Bernard G. Greenberg. 1961. The Ecology of Medical Care. *New England Journal of Medicine* 265, 18 (1961), 885–892. https://doi.org/10.1056/NEJM196111022651805 arXiv:https://doi.org/10.1056/NEJM196111022651805 PMID: 14006536.
[98] David R. Williams and Selina A. Mohammed. 2008. Discrimination and racial disparities in health: Evidence and needed research. *Journal of Behavioral Medicine* 32, 1 (2008), 20–47. https://doi.org/10.1007/s10865-008-9185-0
[99] Yishu Xue, Ofer Harel, and Robert Aseltine. 2019. Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*. IEEE, Bordeaux, France, 1–4. https://doi.org/10.1109/SampTA45681.2019.9030977
[100] Stephanie Yom and Maichou Lor. 2021. Advancing health disparities research: the need to include Asian American subgroup populations. *Journal of Racial and Ethnic Health Disparities* 9 (2021), 1–35. Issue 6. https://pubmed.ncbi.nlm.nih.gov/34791615/
[101] Qianyu Yuan, Tianrun Cai, Chuan Hong, Mulong Du, Bruce E. Johnson, Michael Lanuti, Tianxi Cai, and David C. Christiani. 2021. Performance of a Machine Learning Algorithm Using Electronic Health Record Data to Identify and Estimate Survival in a Longitudinal Cohort of Patients With Lung Cancer. *JAMA Network Open* 4, 7 (07 2021), e2114723–e2114723. https://doi.org/10.1001/jamanetworkopen.2021.14723
[102] Guangyu Zhang, Charles E. Rose, Yujia Zhang, Rui Li, Florence C. Lee, Greta Massetti, and Laura E. Adams. 2022. Multiple imputation of missing race and ethnicity in CDC covid-19 case-level surveillance data. *International Journal of Statistics in Medical Research* 11 (2022), 1–11. https://doi.org/10.6000/1929-6029.2022.11.01
[103] Yan Zhang. 2018. Assessing fair lending risks using race/ethnicity proxies. *Management Science* 64, 1 (2018), 178–197.

A ETHICS AND SOCIAL IMPACTS OF OUR WORK

In this section, we address ethical considerations on the imputation of race and the use of the medical dataset. First, we reiterate our understanding of race as a social construct and the importance of measuring it despite the imperfect choices involved. Second, we discuss the practical implications of the study and the risks involved with inferring race using medical data. Lastly, we provide justification for the use of the dataset.

Throughout this study, we emphasize that race is not a fixed biological or genetic concept, but rather a social construct that categorizes people based on historical and social factors. As Lu et al. [60] suggests, race, ethnicity, and ancestry are distinct but overlapping concepts that should be considered separately in medical research. We recognize that racial categories are not static and can change over time. Consequently, it can be challenging to design a precise measure to capture the effects of race and to define what constitutes an "accurate" racial inference. Additionally, the use of racial labels is becoming more complex as they fail to account for the growing number of mixed-race individuals and the increasing diversity within each racial group.

We emphasize the difference between using race to evaluate disparities and using race to make adverse decisions. Race-based decisions pose grave problems, for instance, for content delivery, lending, employment, or housing [11, 78]. Rather, measuring racial disparities is important precisely because disparities can arise from the social construct of race. Imputation allows us to define, quantify, and bring awareness to racial disparities, despite the imperfect nature of demographic measurement [80]. In fact, there are several instances when there are no legitimate alternatives to race imputation for measuring disparities. U.S. Executive Order 13,985 [29], for example, requires federal agencies to conduct racial equity assessments, even though many of these agencies lack records of individual self-reported race. Similarly, racially polarized voting analyses under the Voting Rights Act [88] require measurement of the majority- and minority-race voters in a district, but self-reported race is again often lacking. Under the government settlement with Meta for violations of the Fair Housing Act [30], Meta agreed to use BISG to estimate the race and ethnicity of its users. Beyond legal requirements for disparity assessments, race imputation can also fulfill an urgent need to understand health disparities, where race is often missing, as emphasized by the U.S. Department of Health and Human Services [71]. Critically, the need for imputation as a tool to evaluate disparities when race is so often missing does not imply that it should be used in decision making [44]. Our goal is not to reify race as a concept, but to improve the reliability of associating people with racial and ethnic categories to enable disparity measurement for race-associated conditions, diagnoses, treatment effects, and outcomes that are not necessarily the result of cultural, societal, and individual biases, regardless of reporting categories used. Our work exploring disaggregation precisely speaks to this.

In terms of practical implications, our study shows that name and geography information alone provide strong signals for predicting race, and additional information derived from EHR data provide only incremental improvements. Predicting race using only name and geography information can produce biased estimates [59]. At the same time, we show that although more information could improve the overall imputation performance, there are different trade-offs for different racial groups. The information and imputation methods used should be dependent on data access and quality. For example, studies show that racial minorities are more likely to be under- or misdiagnosed [21, 33, 34, 45, 62, 89] and that the collection of medical information can be disproportionately less accurate for racial minorities as well [73, 85]. The variance in data quality would likely bias the prediction outcomes.

Lastly, we use this dataset solely for research purposes with IRB approval, including Waiver of Informed Consent, Waiver of Assent, and Waiver of HIPAA Authorization, on servers approved for High Risk and Protected Health Information data. The dataset provides a unique setting to investigate redundant encodings and core questions of health disparities (see Appendix D), as it is a comprehensive dataset that covers many under-served patients from rural, low-income, and racial minority populations. The use of race imputation in medical decision making raises much more acute ethical concerns than the use of imputation to enable the *assessment* of health disparities.

B CONSTRUCTION OF RACE AND ETHNICITY VARIABLE IN THE AMERICAN FAMILY COHORT

B.1 Construction of race variable

The American Family Cohort (AFC) dataset contains two race fields for each patient: *patientracecode* and *patientracetext. patientracecode* contains HL7 codes established by the U.S. Centers for Disease Control and Prevention (CDC). The coding system has six racial categories: American Indian and Alaska Native, Asian, Black or African American, Native Hawaiian and Other Pacific Islander, White, and Other Race. Each category has one parent code (*e.g.*, 1002-5 for American Indian or Alaska Native), and most have multiple child codes (*e.g.*, 1010-8 for Apache), each paired with a text description. *patientracecode* may contain one or more parent codes and/or child codes, which may not match official HL7 codes, or no codes at all. *patientracetext* contains free text. AFC exhibits wide heterogeneity in the consistency and clarity of provider data standards, ranging from codes and text that always exactly match the CDC system to providing no code and only single characters.

A derived race variable was constructed using the following steps. First, for each of the six racial categories, a string match was sought between any code in *patientracecode* and any HL7 code within the category, and between any subset of terms of *patientracetext* and any text description within the category, including some spelling variations. The terms “caucasian”, “wh”, and “w”, among others, were considered text matches to White. The terms “blk”, “b”, and “aa”, among others, were considered text matches to Black or African American. Second, patients were assigned a racial category based on code matches and another based on text matches, where matches in more than one category were assigned Multiple Races. Third, a final racial category was assigned, prioritizing the code-based match over the text-based match. If *patientracecode* matched 2131-1 for Other Race, but *patientracetext* matched to a racial category, the text-based match was used. The remaining patients, who have neither a

code-based or text-based match, including text descriptions such as "unknown", "decline", or "refused", were assigned Unknown.

B.2 Construction of ethnicity variable

The AFC dataset contains two ethnicity fields for each patient: *patientethnicitycode* and *patientethnicitytext*. The CDC HL7 coding system has one parent code and about 40 child codes for Hispanic or Latino and one parent code for Not Hispanic or Latino. A single ethnicity variable was constructed using the same three steps. The terms "hispanic", "latino", "hispa", "his", or "h", among others, were considered text matches to Hispanic or Latino. Any code-based match, or text-based match if there was no code-based match, to Hispanic or Latino led to the final assignment of Hispanic or Latino. All other patients were assigned Not Hispanic or Latino, or Unknown.

B.3 Construction of variable combining race and ethnicity

A final variable combining race and ethnicity prioritizes ethnicity over race. If the ethnicity variable was Hispanic or Latino, then the combined variable was assigned Hispanic or Latino. If the ethnicity variable was Not Hispanic or Latino, or Unknown, then the combined variable was assigned the value of the race variable. If the ethnicity variable was Not Hispanic or Latino, or Unknown, and the race variable was Unknown, then the combined variable was assigned Unknown. In conclusion, the combined variable standardizes across the varied provider information from four AFC data fields into a race and ethnicity classification that aligns with the OMB's 1997 Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity.

C FEATURE CLEANING AND CODING

A detailed description of feature cleaning for each category of features is outlined below. The full list of features can be found at https://github.com/reglab/redundant-encodings.

C.1 Baseline

Baseline features comprise prior probabilities for race given first name, surname given race, and race given geographic area, based on name and geography tables from [14, 92, 93]. Prior to lookup in the name tables, we convert all first and last names to uppercase and keep only the characters before the first whitespace. Because only letter characters A-Z are present in the first name and surname tables, non-letter characters are also removed. To handle patients with duplicate first or last names listed due to typos (*e.g.*, ATHENA v. ATEHNA), we keep the name that matches a name in the Census name tables where possible. For duplicates still unresolved, we keep the patient's given name (name when marital status is Single, then Divorced, then Married), and if still unresolved, we keep the earliest appearance of the patient's name.

For geography priors, we retain only one address for each patient, consisting of street address, city, and ZIP Code, and geocode using a combination of the Census Geocoder and the Google Geocoding API. The geocoded coordinates are then spatially joined to 2020 census block group shapefiles. In the event a census block group-based geography prior cannot be obtained, if the patient has both first and last name priors, our BIFSG algorithm falls back on only those two priors. If the patient only has a last name prior, our BIFSG algorithm uses that in combination with a state-based geography prior.

C.2 Demographic

Age is calculated using the patient's date of birth with respect to the most recent visit date. Gender, marital status, and state are one-hot encoded. A separate missingness feature is created for state; for gender and marital status, we use the "Other/Unknown" category to represent both missingness and other levels.

C.3 Behavioral

Behavioral features are identified by searching free-text patient records of social history observations. Smoking and tobacco history is determined by searching for the terms "smoke", "smok", "smk", and "tobacco" that appear with an affirmative observation, indicated by the terms "yes", "Y", "current", "every day", "some day", and "former", and that do not appear with a negative or unknown observation, indicated by the terms "never", "non", or "unknown".

Language preferences are identified by searching free-text language text and language code fields of patient language records. We limit the set of languages to the most prevalent ones in the dataset with at least 1,000 subjects: English, Spanish, Vietnamese, Chinese, Tagalog, Yiddish, and Armenian. Language preferences are determined by the terms "English", "eng", "en", "151", and "593" for English; "Spanish", "spa", "17117", and "312741" for Spanish; "Vietnamese", "vie", "312743", and "132317644" for Vietnamese; "Chinese", "chi", "zho", "312733", and "17110" for Chinese; "Tagalog" and "tgl" for Tagalog; "Yiddish" and "yid" for Yiddish; and "Armenian", "hye", and "arm" for Armenian.

C.4 Administrative

To extract insurance information, we search several fields including insurance plan, insurance company, documentation date, expiration date, and active or inactive status. We search for terms "blue cross", "blue shield", "bc/bs", "medicare", "medical", "medica", "aetna", "united", "united health care", "united healthcare", "UHC", "cigna", "private", "commercial", "preferred provider organization", "group policy", "hmo/managed care", "humana", "private", "self pay", "capitated", "tricare", and "coventry". Each insurance plan is represented with a Boolean variable, which indicates if a patient has ever had the plan at any time in their record. The total number of recorded insurance plans and the total number of active insurance plans are represented as numeric features.

Visit counts per year are generated by counting the number of unique encounter dates in each year from 2010 to 2022.

C.5 General Health

Allergy features are determined by searching for the following terms in allergy descriptions: "NKDA", "NKA", "no known allergies", "no known medication allergies", "no known drug allergies", "no known active allergies", "penicillin", "sulfa", "sulfonamide", "sulfa drug", "sulfamethoxazole", "latex", "codeine", "aspirin", "pril",

"ACE inhibitor", "angiotensin converting enzyme inhibitor", "doxycycline", "amoxicillin", "statin", "zocor", "morphine", "ibuprofen", "NSAIDs", "iodine", "biaxin", and "prednisone".

Immunization features are generated from immunization records for patients. Because immunization records are available for less than 10% of the patient population, we consider only the count of immunizations a patient has on record and forgo any more granular immunization features.

Vital signs are generated from observation names and codes for body mass index, systolic and diastolic blood pressure, height, weight, heart rate, oxygen saturation, respiratory rate, and temperature. We identify each measurement with both textual terms (*e.g.*, "systolicbp", "systolic", "bpsystolic", "bloodpressuresystolic", and more for systolic blood pressure), as well as Logical Observation Identifiers Names and Codes (LOINC) codes, which provide a clinical standard for laboratory and clinical test results, and SNOMED CT codes (*e.g.*, "8480-6", "271649006", and more for systolic blood pressure). For each measurement, we convert all records to the same units.

C.6 Medical

Diagnosis codes are generated from current and historical clinical problems. We identify the type of code (ICD-10, ICD-9, or SNOMED CT) by the listed problem category, or with regular expression patterns for codes with an ambiguous category. ICD-9 and SNOMED CT codes are mapped to their ICD-10 equivalent with crosswalks published by [66] and [67], respectively. After performing a χ^2 test and calculating the mutual information between diagnosis codes and race class labels, we take the union of codes from the literature described in Appendix D and codes with non-zero mutual information and $p \leq 0.05$. Of these codes, the 1,000 codes with the highest mutual information are selected.

Procedure codes are documented in the Current Procedural Terminology (CPT) and the Healthcare Common Procedure Coding System (HCPCS). These are collections of standardized codes that represent medical procedures, supplies, products, and services. All codes are aggregated at the CPT Category I and HCPCS Level II levels of aggregation, which include high-level categorization of procedures such as evaluation & management, anesthesia, surgery, radiology procedures, pathology, and laboratory procedures. We additionally include more granular CPT Category II codes. The final feature set contains 55 categories of procedures and the total counts of procedures for each patient.

D LITERATURE REVIEW FOR SELECTING DIAGNOSIS CODES

There are over 68,000 ICD-10 codes [10], which poses computational challenges in feature encoding, hyperparameter searching, and model training. To filter the full set of diagnostic codes to the most relevant ones, we conduct a substantive literature review and include diagnosis codes that appear in the literature to disproportionately impact certain racial groups. Table 5 shows examples of these ICD-10 codes, with their prevalence by racial group (per 10,000 patients) for our study population in the AFC dataset. For racial minority groups, we additionally indicate with an asterisk which codes are documented in the literature to have a higher prevalence compared to the White population. We observe differences in prevalence among racial groups that largely comport with existing literature, though we note that the AFC data provides unique opportunities for future work to reexamine and expand on existing knowledge about racial disparities in health through its rich and traditionally underrepresented cohort.

E HYPERPARAMETER TUNING PROCEDURE

We use the following training procedure to select optimal hyperparameters and train and test each random forest model. We first randomly split the study population into training and test sets representing 80% and 20% of the data, respectively. The train/test split is stratified on the race class label to preserve the percentage of each racial group in each split.

Next, we tune the random forest hyperparameters to select optimal values. To do so, we perform five-fold cross validation within the 80% training set, splitting the training data into five folds, and then iterating through each fold, using one fold as the validation set and training on the remaining folds. We use this five-fold cross validation procedure to perform a search over the following distribution of parameters:

- Number of estimators: 100, 1000
- Maximum depth of tree: 3, 10, 15, 20, 25, 50, None
- Minimum samples required for a leaf node: 1, 10, 25, 50
- Minimum samples required to split a node: 2, 10, 25, 50, 100
- Number of features to consider: 25%, 50%, 75%, square root of total number of features

We perform a randomized search (as opposed to a grid search) due to the magnitude of the data, with over 4 million training examples, which leads to long computation times to perform the five-fold cross validation over a large number of hyperparameter combinations. All random forest models use bootstrapped samples when building the trees. We use the same train/test split across all random forest models, but perform a separate hyperparameter search for each feature set we consider (*i.e.*, baseline, demographic, etc.).

We select the hyperparameter values from the search with the highest mean micro AUPRC across the five validation folds, with a percent difference in train and validation performance less than 2% to prevent overfitting. We choose AUPRC instead of the typical AUROC because of the highly imbalanced nature of the dataset, and because we wish to make positive predictions while minimizing false positive predictions for the minority groups. This is also aligned with the conventional practice. (We are exploring a cost-sensitive learning approach for a future iteration of the work, which would allow us to explicitly increase the cost of mis-classifying a minority group member and control the level of disparity among groups.) We evaluate at the micro level so we can choose the model that performs the best at the individual level. This also means that the model does not explicitly consider the group disparity. In the main text we present and contextualize both micro and macro level statistics. Using this set of optimal hyperparameter values, we train the final model on the 80% training set. We evaluate the trained model on the unseen 20% test set.

Table 5: Example literature-based diagnosis codes. For each code in the literature, we show the prevalence by racial group (per 10,000 patients) for our study population. An asterisk indicates for which racial minority groups the codes are documented in the literature to have a higher prevalence compared to the White population.

Description	*ICD-10 Code*	*AIAN*	*ASIAN*	*BLACK*	*HISP*	*NHPI*	*WHITE*
Asthma [70]	J45	900*	700	1,000*	830	1,300*	820
Stomach cancer [70]	C16	3.7*	7.5*	5.3*	3.7*	2.5*	3
Liver & IBD cancer [70]	C22	3.7*	7.5*	5.3*	5.6*	12*	5.2
Diabetes [65, 70]	E08-E13	1,400*	1,500*	1,900*	1,500*	1,500*	1,300
Alcohol use disorder [31]	F10	170*	41	100	97	110	140
Illicit drug use disorder [31]	F11-F16, F18-F19	210*	43	130	110	160	150
Coronary heart disease [70]	I20-I25	610*	360	460	360	440*	700
Chronic hepatitis B [70]	B16	1.7	23*	6	2.9	9.1*	2.6
Underimmunization [70]	Z28	150*	260*	320*	270*	300	230
Perinatal conditions [70]	P00-P96	320*	170	350*	290*	150*	250
History of pyschological trauma [7, 70, 98]	Z91.4	5.4*	0.86	4.4	3.6*	2.5	3.1
Obesity [65, 70]	E65-E68	1,300*	830	2,000*	1,600*	1,700*	1,400
Transplanted organ [70]	Z94	9.1	11	16*	12	11	11
End stage renal disease [33]	N18.6	43*	23	73*	37	45	15
Sickle cell disorders [58, 76]	D57	11	2.9	67*	4.7	6.6	3.0
Homelessness [5, 63]	Z59.0	5.8*	0.69	7.1*	2.7*	2.5	4.7
Syphilis [53]	A50-A53	8.3	7.4	33*	16	6.6	10
Tuberculosis [53]	A15-A19	5*	25*	8*	7.5*	11*	3.5

F ADDITIONAL METRICS

Table 6 provides AUPRC, AUROC, F1-score, precision, and recall for each racial group for the conventional BIFSG and ML BIFSG models; the use of ML leads to an increase in imputation performance for each of these metrics for every racial group. Table 7 shows micro and macro AUPRC, AUROC, F1-score, precision, and recall for an increasing feature set, beginning with baseline features alone, and incrementally growing the feature set. Table 8 shows AUPRC, AUROC, F1-score, precision, and recall for each racial group with an increasing feature set. 95% confidence intervals are shown for each of these tables.

We also further explore the risk of redundant encodings for features beyond name and geography. As described in §4.2, the risk of redundant encodings is not exacerbated with the addition of our richer feature sets beyond baseline features. To provide supplementary evidence of this finding, we compare the performance of five random forest models with increasing feature sets, beginning with demographic features and incrementally adding behavioral, administrative, general health, and medical features; all models exclude baseline features. These models represent a realistic setting when name or geography information may be unavailable. Figure 4 displays the micro and macro AUPRC and AUROC for the five random forest models; the x-axis shows the feature set on which each of the five models was trained. The performance for each individual feature set is shown with "x" markers. Though micro and macro AUPRC and AUROC all increase monotonically from demographic features alone to the full feature set excluding baseline features, the performance never reaches that of baseline features alone. This finding provides further evidence that, although additional features do encode further probabilistic information about race, concerns of redundant encodings are most warranted when name and geography features are readily accessible.

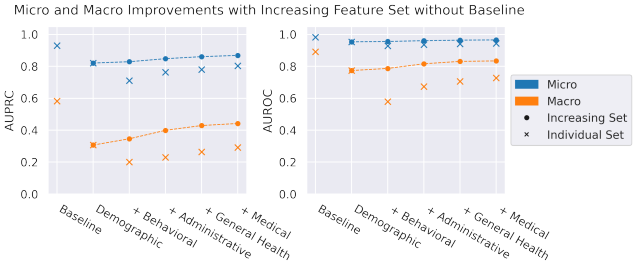


Figure 4: Micro and macro performance for an increasing feature set with baseline features excluded, with AUPRC on the left and AUROC on the right. "x" markers indicate the micro and macro performance for each feature set alone. Error bars representing 95% confidence intervals are too tight to be seen. In situations where only anonymized data is available to researchers, we observe improvement in both micro and macro AUPRC and AUROC when adding more features. However, neither the model trained with the largest feature set nor the models trained on each of the individual feature sets perform as well as the model trained with baseline features.

G INVESTIGATION OF MECHANISM BY WHICH ML IMPROVES UPON BI(F)SG

To better understand the mechanism through which ML improves upon conventional BIFSG, we train two Logistic Regression (LR) models with BIFSG priors, and with BIFSG priors plus the interaction terms of the priors, respectively. By including the interaction terms, we explicitly provide hints that there may be interactions between race and geographic area (*e.g.*, the same name may be associated with different probabilities of being a certain race depending on the geographic location) to the otherwise linear model. We would expect to observe improvement in performance upon LR with only BIFSG priors by including interaction terms if the independence assumption of BIFSG is violated in practice. We would

Table 6: Machine learning improvements to BIFSG by race with comparable feature set, with 95% confidence intervals.

	Model	*AIAN*	*API*	*BLACK*	*HISP*	*WHITE*
AUPRC	Conventional BIFSG	0.048 (0.041-0.054)	0.694 (0.688-0.699)	0.552 (0.548-0.556)	0.738 (0.735-0.741)	0.953 (0.953-0.954)
	ML BIFSG	0.124 (0.113-0.134)	0.759 (0.754-0.764)	0.629 (0.626-0.632)	0.853 (0.851-0.855)	0.964 (0.964-0.964)
AUROC	Conventional BIFSG	0.690 (0.682-0.696)	0.909 (0.906-0.911)	0.879 (0.878-0.880)	0.933 (0.932-0.933)	0.896 (0.895-0.896)
	ML BIFSG	0.832 (0.825-0.837)	0.941 (0.938-0.943)	0.917 (0.917-0.918)	0.954 (0.954-0.955)	0.917 (0.917-0.918)
F1-Score	Conventional BIFSG	0.085 (0.076-0.096)	0.706 (0.702-0.710)	0.472 (0.469-0.476)	0.798 (0.796-0.799)	0.918 (0.918-0.919)
	ML BIFSG	0.070 (0.059-0.078)	0.751 (0.747-0.755)	0.521 (0.518-0.524)	0.829 (0.828-0.831)	0.928 (0.927-0.928)
Precision	Conventional BIFSG	0.241 (0.219-0.272)	0.821 (0.816-0.826)	0.647 (0.642-0.651)	0.785 (0.783-0.787)	0.893 (0.893-0.894)
	ML BIFSG	0.754 (0.698-0.814)	0.845 (0.841-0.851)	0.742 (0.740-0.746)	0.823 (0.821-0.825)	0.900 (0.899-0.900)
Recall	Conventional BIFSG	0.051 (0.046-0.059)	0.619 (0.615-0.624)	0.372 (0.369-0.375)	0.812 (0.810-0.814)	0.945 (0.944-0.945)
	ML BIFSG	0.037 (0.031-0.041)	0.676 (0.670-0.680)	0.401 (0.398-0.405)	0.836 (0.834-0.838)	0.958 (0.957-0.958)

Table 7: Micro and macro performance for an increasing feature set, with 95% confidence intervals.

Model	*AUPRC*	*AUROC*	*F1-Score*	*Precision*	*Recall*
			Micro		
Baseline	0.933 (0.933-0.934)	0.983 (0.983-0.983)	0.881 (0.880-0.881)	0.881 (0.880-0.881)	0.881 (0.880-0.881)
+ Demographic	0.939 (0.938-0.939)	0.984 (0.984-0.985)	0.886 (0.886-0.887)	0.886 (0.886-0.887)	0.886 (0.886-0.887)
+ Behavioral	0.940 (0.940-0.941)	0.985 (0.985-0.985)	0.887 (0.887-0.888)	0.887 (0.887-0.888)	0.887 (0.887-0.888)
+ Administrative	0.942 (0.941-0.942)	0.985 (0.985-0.985)	0.889 (0.889-0.890)	0.889 (0.889-0.890)	0.889 (0.889-0.890)
+ General Health	0.943 (0.943-0.944)	0.985 (0.985-0.986)	0.891 (0.890-0.891)	0.891 (0.890-0.891)	0.891 (0.890-0.891)
+ Medical	0.944 (0.943-0.944)	0.985 (0.985-0.986)	0.892 (0.892-0.892)	0.892 (0.892-0.892)	0.892 (0.892-0.892)
			Macro		
Baseline	0.585 (0.581-0.588)	0.893 (0.891-0.896)	0.541 (0.537-0.544)	0.782 (0.768-0.796)	0.504 (0.501-0.506)
+ Demographic	0.598 (0.594-0.602)	0.901 (0.899-0.903)	0.549 (0.545-0.553)	0.792 (0.777-0.803)	0.509 (0.506-0.511)
+ Behavioral	0.602 (0.598-0.605)	0.904 (0.902-0.905)	0.551 (0.547-0.554)	0.794 (0.779-0.806)	0.510 (0.507-0.512)
+ Administrative	0.610 (0.607-0.614)	0.906 (0.904-0.908)	0.557 (0.554-0.561)	0.806 (0.793-0.815)	0.514 (0.512-0.516)
+ General Health	0.618 (0.614-0.621)	0.908 (0.906-0.910)	0.562 (0.558-0.565)	0.824 (0.811-0.833)	0.517 (0.515-0.519)
+ Medical	0.621 (0.618-0.625)	0.907 (0.906-0.909)	0.567 (0.563-0.570)	0.822 (0.809-0.831)	0.520 (0.518-0.522)

Table 8: Performance by race for an increasing feature set, with 95% confidence intervals.

	Model	*AIAN*	*ASIAN*	*BLACK*	*HISP*	*NHPI*	*WHITE*
AUPRC	Baseline	0.124 (0.113-0.134)	0.785 (0.780-0.790)	0.629 (0.626-0.632)	0.853 (0.851-0.855)	0.152 (0.134-0.169)	0.964 (0.964-0.964)
	+ Demographic	0.135 (0.125-0.146)	0.795 (0.790-0.800)	0.654 (0.651-0.657)	0.873 (0.871-0.875)	0.164 (0.143-0.181)	0.968 (0.967-0.968)
	+ Behavioral	0.143 (0.132-0.152)	0.799 (0.794-0.804)	0.656 (0.654-0.660)	0.880 (0.879-0.882)	0.165 (0.145-0.182)	0.968 (0.968-0.968)
	+ Administrative	0.174 (0.162-0.184)	0.800 (0.795-0.805)	0.662 (0.660-0.666)	0.886 (0.885-0.888)	0.171 (0.152-0.187)	0.969 (0.968-0.969)
	+ General Health	0.198 (0.185-0.207)	0.806 (0.801-0.811)	0.670 (0.668-0.674)	0.891 (0.889-0.892)	0.175 (0.157-0.192)	0.970 (0.969-0.970)
	+ Medical	0.198 (0.186-0.210)	0.806 (0.801-0.811)	0.681 (0.679-0.685)	0.891 (0.890-0.892)	0.182 (0.164-0.200)	0.970 (0.970-0.970)
AUROC	Baseline	0.832 (0.825-0.837)	0.956 (0.954-0.957)	0.917 (0.917-0.918)	0.954 (0.954-0.955)	0.783 (0.769-0.794)	0.917 (0.917-0.918)
	+ Demographic	0.838 (0.831-0.844)	0.962 (0.960-0.964)	0.925 (0.925-0.926)	0.959 (0.959-0.960)	0.797 (0.786-0.807)	0.924 (0.923-0.925)
	+ Behavioral	0.841 (0.834-0.846)	0.962 (0.960-0.964)	0.926 (0.925-0.927)	0.961 (0.960-0.961)	0.807 (0.797-0.817)	0.925 (0.925-0.926)
	+ Administrative	0.843 (0.836-0.848)	0.962 (0.961-0.964)	0.927 (0.926-0.928)	0.961 (0.961-0.962)	0.816 (0.806-0.825)	0.927 (0.926-0.927)
	+ General Health	0.845 (0.838-0.850)	0.965 (0.964-0.967)	0.929 (0.928-0.930)	0.962 (0.962-0.963)	0.816 (0.803-0.826)	0.929 (0.928-0.929)
	+ Medical	0.839 (0.833-0.845)	0.964 (0.963-0.966)	0.931 (0.930-0.932)	0.962 (0.961-0.963)	0.817 (0.805-0.827)	0.930 (0.929-0.930)
F1-Score	Baseline	0.070 (0.059-0.078)	0.772 (0.767-0.776)	0.521 (0.518-0.524)	0.829 (0.828-0.831)	0.126 (0.108-0.143)	0.928 (0.927-0.928)
	+ Demographic	0.070 (0.059-0.079)	0.774 (0.770-0.779)	0.544 (0.542-0.548)	0.843 (0.841-0.844)	0.132 (0.114-0.148)	0.931 (0.931-0.931)
	+ Behavioral	0.071 (0.060-0.079)	0.779 (0.775-0.783)	0.547 (0.544-0.550)	0.845 (0.843-0.846)	0.131 (0.113-0.146)	0.932 (0.931-0.932)
	+ Administrative	0.102 (0.089-0.112)	0.779 (0.775-0.783)	0.554 (0.551-0.557)	0.849 (0.848-0.851)	0.128 (0.110-0.144)	0.933 (0.932-0.933)
	+ General Health	0.123 (0.110-0.131)	0.781 (0.777-0.785)	0.558 (0.555-0.561)	0.851 (0.850-0.852)	0.127 (0.107-0.142)	0.933 (0.933-0.934)
	+ Medical	0.134 (0.122-0.144)	0.781 (0.777-0.785)	0.566 (0.563-0.569)	0.852 (0.851-0.853)	0.133 (0.113-0.150)	0.934 (0.934-0.935)
Precision	Baseline	0.754 (0.698-0.814)	0.829 (0.824-0.834)	0.742 (0.740-0.746)	0.823 (0.821-0.825)	0.645 (0.582-0.692)	0.900 (0.899-0.900)
	+ Demographic	0.771 (0.717-0.832)	0.835 (0.829-0.840)	0.756 (0.753-0.759)	0.854 (0.852-0.855)	0.635 (0.565-0.682)	0.901 (0.900-0.901)
	+ Behavioral	0.757 (0.704-0.820)	0.836 (0.831-0.841)	0.759 (0.757-0.763)	0.862 (0.860-0.863)	0.646 (0.579-0.688)	0.901 (0.900-0.901)
	+ Administrative	0.811 (0.775-0.853)	0.837 (0.832-0.843)	0.766 (0.763-0.769)	0.867 (0.865-0.868)	0.650 (0.594-0.694)	0.902 (0.901-0.902)
	+ General Health	0.858 (0.828-0.897)	0.840 (0.835-0.846)	0.774 (0.771-0.778)	0.869 (0.868-0.871)	0.698 (0.629-0.740)	0.902 (0.902-0.903)
	+ Medical	0.865 (0.831-0.895)	0.841 (0.836-0.846)	0.788 (0.784-0.791)	0.870 (0.869-0.872)	0.665 (0.602-0.714)	0.903 (0.902-0.904)
Recall	Baseline	0.037 (0.031-0.041)	0.722 (0.716-0.727)	0.401 (0.398-0.405)	0.836 (0.834-0.838)	0.070 (0.059-0.080)	0.958 (0.957-0.958)
	+ Demographic	0.037 (0.031-0.041)	0.722 (0.716-0.727)	0.425 (0.423-0.429)	0.832 (0.830-0.833)	0.074 (0.063-0.083)	0.963 (0.963-0.963)
	+ Behavioral	0.037 (0.031-0.042)	0.729 (0.723-0.734)	0.427 (0.424-0.430)	0.828 (0.826-0.830)	0.073 (0.062-0.082)	0.965 (0.964-0.965)
	+ Administrative	0.054 (0.047-0.060)	0.728 (0.722-0.733)	0.434 (0.430-0.437)	0.833 (0.831-0.834)	0.071 (0.061-0.081)	0.966 (0.965-0.966)
	+ General Health	0.066 (0.059-0.071)	0.730 (0.725-0.735)	0.436 (0.433-0.439)	0.833 (0.832-0.835)	0.070 (0.058-0.079)	0.967 (0.966-0.967)
	+ Medical	0.073 (0.066-0.079)	0.729 (0.723-0.734)	0.442 (0.439-0.445)	0.834 (0.832-0.835)	0.074 (0.062-0.084)	0.968 (0.967-0.968)

also expect the performance of the LR with interaction term to be on par with BIFSG if the main mechanism through which ML models improve upon BIFSG is by accounting for more complex interaction among the features.

According to Table 9, we observe small improvement in all measures when comparing LR with interaction terms with the LR with only BIFSG priors. However, when looking at macro performance, LR (BIFSG + Interaction) does not perform better than the conventional BIFSG on AUPRC, AUROC, and F1-score. It also performs uniformly worse than the Random Forest that only uses BIFSG features. Together this shows that the ML Random Forest model does not improve upon conventional BIFSG by simply adjusting for the difference in base rate between AFC population and the census population or by being able to consider interaction terms and account for the violation of independence assumptions. ML likely performs better than conventional BIFSG by being able to identify non-linear relationships among features.

To further identify the relationship among the features, we take one subtree from the random forest, and apply the first three splitting rules to obtain a smaller subsample that can nonetheless demonstrate some of the nonlinearities based on the decision rules. In Figure 5, we show, for each individual, given their last name prior of being a certain race, the probability they would be predicted to be a particular race. We observe non-linear relationships that cannot be captured by a simple linear model.

H DISAGGREGATION OF THE API CATEGORY TO UNDERSTAND HEALTH DISPARITIES

To illustrate how ML-based imputation methods can capture underlying racial health disparities when race is missing, we compute the prevalence of asthma and obesity using the conventional BIFSG technique, ML-based BIFSG, and the ML imputation model that includes all features except medical ones (the asthma and obesity diagnosis codes are included in the medical features). Table 10 shows the prevalence of each condition using a weighting estimator, weighting each positive diagnosis by the subject's predicted race probabilities (*e.g.*, a patient with asthma who has a 50% probability of being White and a 50% probability of being API counts as 0.5 asthma prevalence for White and 0.5 prevalence for API). Ground truth is established using patient's self-reported race.

We select asthma and obesity because these conditions have substantial disparities between the Asian and NHPI populations in the AFC dataset and documented in the literature [65, 70]. We can see from ground truth that the NHPI population has a higher prevalence than the Asian population for both asthma and obesity, a disparity hidden by the collapsed API category. Because conventional BIFSG reports only the aggregated API category, the disparities between Asian and NHPI subjects cannot be captured. Turning to the ML-based methods, both ML BIFSG and ML (+ General Health) capture the disparity for both conditions. As we add more features, the accuracy in the level of disparity increases. These findings highlight the importance of disaggregation and the applicability of ML-based imputation approaches to uncover health disparities between subgroups.

Table 9: Micro and macro performance of machine learning methods and BIFSG with comparable feature set, with 95% confidence intervals.

Model	*AUPRC*	*AUROC*	*F1-Score*	*Precision*	*Recall*
			Micro		
Conventional BIFSG	0.898 (0.898-0.899)	0.967 (0.967-0.967)	0.864 (0.864-0.865)	0.864 (0.864-0.865)	0.864 (0.864-0.865)
LR (BIFSG)	0.906 (0.905-0.906)	0.976 (0.976-0.976)	0.856 (0.855-0.856)	0.856 (0.855-0.856)	0.856 (0.855-0.856)
LR (BIFSG + Interaction)	0.908 (0.908-0.909)	0.977 (0.977-0.977)	0.857 (0.856-0.857)	0.857 (0.856-0.857)	0.857 (0.856-0.857)
RF (BIFSG)	0.934 (0.933-0.934)	0.980 (0.980-0.980)	0.881 (0.881-0.881)	0.881 (0.881-0.881)	0.881 (0.881-0.881)
			Macro		
Conventional BIFSG	0.597 (0.595-0.599)	0.861 (0.859-0.863)	0.596 (0.594-0.599)	0.677 (0.673-0.683)	0.560 (0.558-0.562)
LR (BIFSG)	0.508 (0.506-0.511)	0.850 (0.850-0.852)	0.486 (0.482-0.489)	0.707 (0.693-0.718)	0.460 (0.457 - 0.461)
LR (BIFSG + Interaction)	0.515 (0.512-0.518)	0.858 (0.856-0.860)	0.491 (0.487-0.494)	0.707 (0.692-0.720)	0.465 (0.462-0.467)
RF (BIFSG)	0.666 (0.663-0.668)	0.912 (0.911-0.913)	0.620 (0.617-0.621)	0.813 (0.802-0.825)	0.581 (0.580-0.583)

* RF (BIFSG) is the same as "ML BIFSG" in Table 3. We rename it here for a more clear comparison between the methods.

Table 10: Prevalence of asthma and obesity (per 10,000 patients) by API and disaggreated Asian and NHPI groups. AFC dataset ground truth (via self-reported race) is compared to predictions by conventional BIFSG (restricted to API), ML BIFSG, and ML + General Health (via weighted estimator). Weighted means and weighted standard errors (in parentheses) are shown for each imputation model.

Diagnosis	*Race*	*Ground Truth*	*Conventional BIFSG*	*ML BIFSG*	*ML (+ General Health)*
Asthma	API	752.7	620.4 (18.6)	653.9 (18.3)	667.9 (18.4)
	Asian	688.6	–	620.9 (18.1)	635.6 (18.2)
	NHPI	1363.3	–	1587.1 (71.5)	1579.4 (70.7)
Obesity	API	925.5	816.6 (21.2)	818.9 (20.3)	818.2 (20.2)
	Asian	844.6	–	799.1 (20.3)	791.3 (20.1)
	NHPI	1696.9	–	1378.8 (67.4)	1576.8 (70.6)

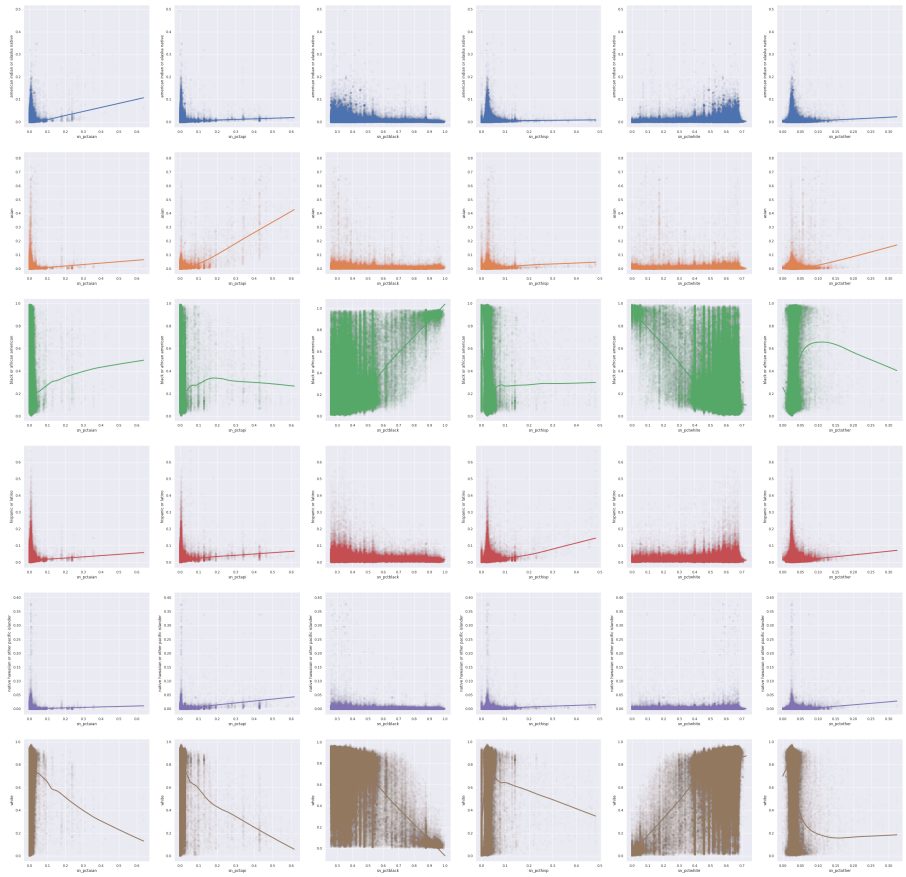


Figure 5: Predicted probabilities of being a race (y-axis, from top to bottom: AIAN, Asian, Black, Hispanic, NHPI, White) given the last name priors of being a race (x-axis, from left to right: AIAN, API, Black, Hispanic, White, Other). We observe some non-linear relationships – for example, between surname priors and the probabilities of being predicted to be Black – which are captured by the RF model.