www.nature.com/scientificdata

scientific data

OPEN

DATA DESCRIPTOR

Enabling disaggregation of Asian American subgroups: a dataset of Wikidata names for disparity estimation

Qiwei Lin[1,7], Derek Ouyang[2,7], Cameron Guage[3], Isabel O. Gallegos[2,4], Jacob Goldin[5] & Daniel E. Ho[2,4,6] ✉

Decades of research and advocacy have underscored the imperative of surfacing – as the first step towards mitigating – racial disparities, including among subgroups historically bundled into aggregated categories. Recent U.S. federal regulations have required increasingly disaggregated race reporting, but major implementation barriers mean that, in practice, reported race data continues to remain inadequate. While imputation methods have enabled disparity assessments in many research and policy settings lacking reported race, the leading name algorithms cannot recover disaggregated categories, given the same lack of disaggregated data from administrative sources to inform algorithm design. Leveraging a Wikidata sample of over 300,000 individuals from six Asian countries, we extract frequencies of 25,876 first names and 18,703 surnames which can be used as proxies for U.S. name-race distributions among six major Asian subgroups: Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese. We show that these data, when combined with public geography-race distributions to predict subgroup membership, outperform existing deterministic name lists in key prediction settings, and enable critical Asian disparity assessments.

Background & Summary

Motivation. Access to high-quality race and ethnicity data is critical for researchers, advocates, and government agencies to understand racial disparities and to design policies that promote racial equity. Race data is often too coarse for meaningful equity assessments[1], and we too often lack the information necessary to identify and act upon existing disparities across diverse populations. In settings with inadequate race data, a range of imputation methods has been developed and deployed to study racial disparities[2], like Bayesian Improved Surname Geocoding (BISG)[3,4]. Such name algorithms are increasingly important and have been used to assess racial disparities across a wide range of settings, including fair lending[5–7], voting[8,9], housing[10], insurance[11], and taxation[12]. However, because of the data with which they are typically implemented, these existing approaches practically restrict imputation to coarse racial categories. Our contribution is to extend BISG to conduct disparity estimation at the Asian subgroup level, responding to recently revised federal standards and long-standing civil rights calls for data disaggregation.

Existing research attempting to impute disaggregated Asian subgroups relies on deterministic name lists sourced from large, restricted datasets[13]; we instead extract name frequencies from over 300,000 individuals in Wikidata and derive proxies for U.S. name-race distributions among Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese subgroups, which can be used as inputs to name algorithms like BISG. Despite using a smaller and less conventional data source, our augmented BISG approach outperforms existing methods in predicting individual race, and performs on par with existing methods in predicting group-level health disparities, for Asian subgroups. Our approach also critically enables probabilistic estimation, facilitating uncertainty

[1]Department of Sociology, Stanford University, Stanford, 94305, USA. [2]Stanford Law School, Stanford University, Stanford, 94305, USA. [3]Department of Economics, Columbia University, New York, 10027, USA. [4]Department of Computer Science, Stanford University, Stanford, 94305, USA. [5]University of Chicago Law School, University of Chicago, Chicago, 60637, USA. [6]Department of Political Science, Stanford University, Stanford, 94305, USA. [7]These authors contributed equally: Qiwei Lin, Derek Ouyang. ✉e-mail: deho@stanford.edu

quantification and control over the precision-recall tradeoff. We release these Wikidata-derived name-race distribution tables along with detailed and flexible guidance on their use within racial imputation algorithms.

Institutional context of race reporting in the U.S. U.S. federal law during the Biden administration mandated racial disparity assessments for agencies and contractors[14,15], and 2024 revisions to federal data collection standards significantly increased the level of disaggregation required in the collection of race and ethnicity information[16]. Notwithstanding the shifting priorities of the proceeding administration, such efforts face major nonpartisan challenges. First, many entities in practice do not collect, and may be affirmatively prohibited under existing laws and regulations from collecting, race information[7,17–19]. Second, as a practical matter, it takes years for federal agencies to implement revisions of such standards. The action plan for the 2024 revisions, for instance, is not set to be finalized until 2029[16]. There is strong evidence, however, to suggest that it would take even longer to achieve full compliance: at present, many agencies have yet to even comply with 1997 standards[20–26]. Third, federal standards take even longer to affect state and private actors. Prior research[25], for instance, documents that few state Medicaid programs collect data that include all minimum categories required by federal standards. Private organizations like insurance companies and hospitals, to the degree they interface with federal programs, may face similar reporting requirements (and challenges) as previously noted[7,27,28], but in general are even less likely to have adequate race information for disparity assessments, given the lack of standardized reporting practices, and other barriers to data collection[17,19,27]. Fourth, such new standards are, of course, only prospective. Records collected or reported prior to revisions (including the 2020 Decennial Census) often practically can never be made fully consistent with the new standards. To the extent that our understanding of racial disparities today is dependent on understanding trends and insights from *historical* data, we cannot rely on new federal standards alone to unlock longitudinal racial equity assessments. Finally, because the *required disaggregation* of reporting options by agencies does not translate to the *required reporting* of disaggregated options by individuals, even absent all the aforementioned challenges, low response rates may yield underpowered and biased samples[29]. Response hesitancy may be particularly salient for certain subgroups given variations in language proficiency[23,30] and concerns about privacy and discrimination[31].

Calls for Asian disaggregation. While inadequate race reporting affects all racial disparity assessments, it is particularly acute for those within the "Asian American" community. The 1977 federal standards defined the single category of Asian or Pacific Islander (API) in its minimum reporting requirement, alongside White, Black, Hispanic, American Indian and Alaska Native (AIAN), and Other. Although the widely employed 1997 revised standards separated the API category into *(1)* Asian and *(2)* Native Hawaiian or Other Pacific Islander (NHPI) categories[20,32], reporting at the Asian level still masks important within-group differences[33–35]. Civil rights advocates note that the term "Asian American," for instance, is a social construct developed in the 1960s[36], which in fact encompasses a wide heterogeneity of subgroups with different countries of origin, language proficiency, socioeconomic status, migration profiles, and baseline health conditions[22]. The lack of disaggregated data and the so-called model minority myth[37,38], which conceives of Asian Americans as a more homogeneous, high-achieving group, can obfuscate within-group disparities.

In the limited cases where disaggregated Asian race data are available for research, studies have identified a range of staggering within-group socioeconomic disparities among Asian Americans in income[39], food insecurity[40], English proficiency, poverty rates, educational attainment[41], and more. In the healthcare setting, research comparing federal data on the life expectancy of Asian versus White populations may yield the headline finding that Asian Americans outlive White Americans by eight years[42]; however, research that takes the additional step of disaggregating the same federal data into separate Asian subgroups unmasks even larger disparities, such as a Vietnamese lag of 9.3 to 11.6 years behind their Chinese counterparts[43,44]. Other public health and clinical research has revealed salient within-group disparities in obesity[45], diabetes[46], and cancer[47], but significant gaps in our understanding of Asian health disparities remain[48].

Limitations of existing approaches. In the absence of reported race data, most studies on within-group disparity among Asian Americans rely on Asian name lists to make categorical classifications[49–51]. The most commonly used name lists[13], published in 2000, are sourced from Social Security Administration (SSA) files and are not widely available for public use. Devoid of any frequency information, the name lists instead deterministically link names to subgroups via a one-to-one mapping (*e.g.*, Sato to Japanese, Singh to Asian Indian). However, more recent work has sought to better account for the uncertainty in the relationship between name and race through the development of probabilistic methods such as BISG. These methods, which use Bayesian updating to combine information about the distribution of race across names and geographic areas to make probabilistic predictions about a person's race, have been used across a variety of consequential domains[5–12], and offer several key advantages over deterministic name lists. First, name algorithms assign varying probabilities of belonging to different racial groups instead of assigning discrete labels. Validation studies offer evidence that using categorical classification or discretizing predictions into classification labels often leads to undercounting of minority groups and downstream inference bias[52,53]. By not discretizing, name algorithms enable researchers to utilize estimated race probabilities within estimators that require more than a categorical classification[12,54,55]. Researchers can also choose a specific classification threshold (*i.e.*, a 90% probability or higher of being a certain race) for their particular prediction setting. This is highly consequential, as the ideal threshold – and the trade-off between precision and recall – may vary substantially across tasks[53,56]. In contrast, name lists lose information because they do not differentiate the varying predictive power of names along a continuous range, with some being more racially distinctive than others, and they exclude names with only moderate specificity[3]. Second, name list methods do not specify how to adjudicate conflicting classifications when first and last names are associated with different groups, so if an individual has a first name and last name yielding divergent classifications, researchers can only rely on one of the lists to make a decision. Name algorithms, on the other hand, use Bayesian updating to incorporate

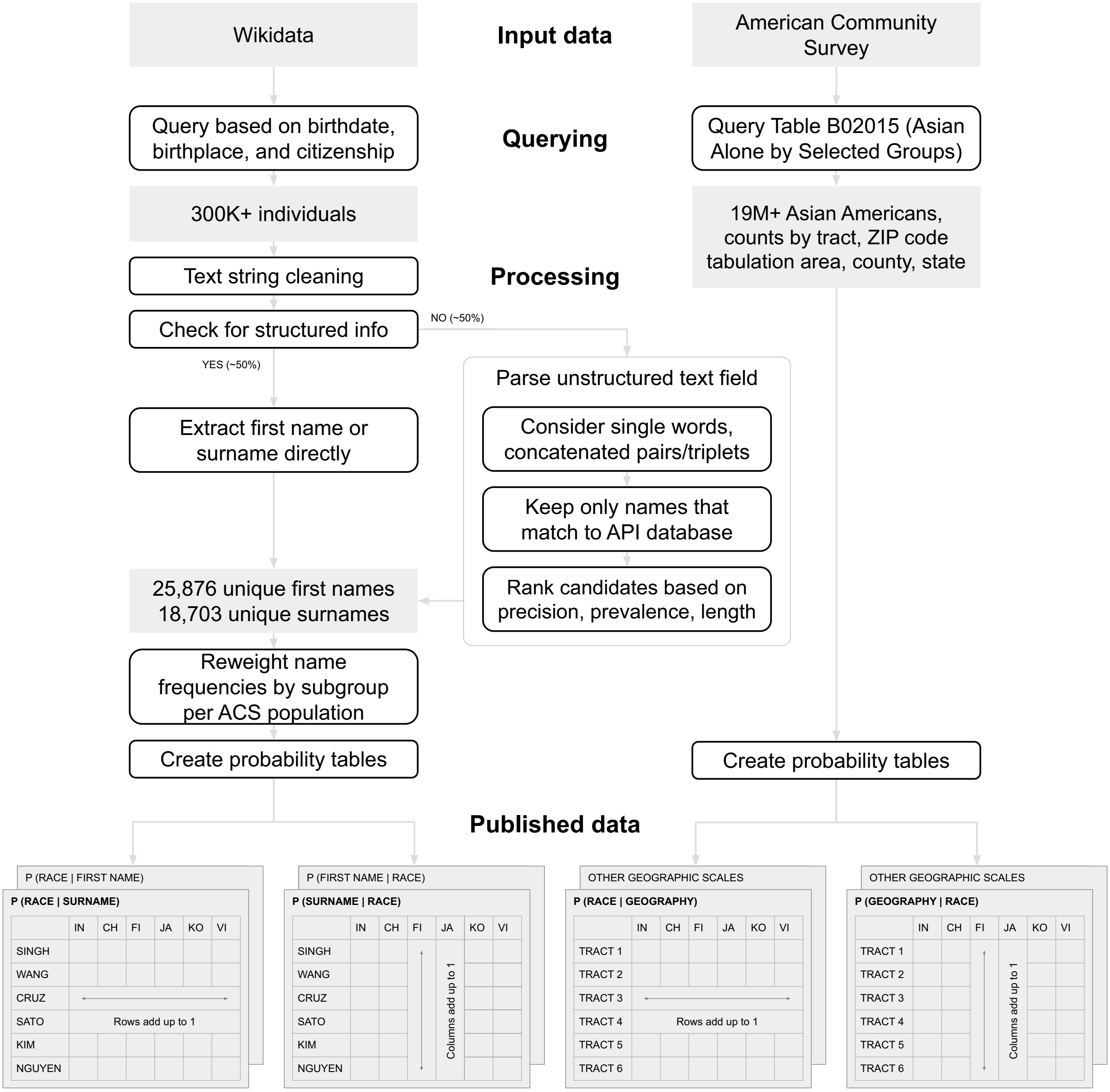
Wikidata
Input data
American Community Survey
Query based on birthdate, birthplace, and citizenship
Querying
Query Table B02015 (Asian Alone by Selected Groups)
300K+ individuals
19M+ Asian Americans, counts by tract, ZIP code tabulation area, county, state
Text string cleaning
Processing
Check for structured info
NO (~50%)
YES (~50%)
Parse unstructured text field
Consider single words, concatenated pairs/triplets
Keep only names that match to API database
Rank candidates based on precision, prevalence, length
Extract first name or surname directly
25,876 unique first names
18,703 unique surnames
Reweight name frequencies by subgroup per ACS population
Create probability tables
Create probability tables
Published data
P (RACE | FIRST NAME)
P (RACE | SURNAME)
P (FIRST NAME | RACE)
P (SURNAME | RACE)
OTHER GEOGRAPHIC SCALES
P (RACE | GEOGRAPHY)
OTHER GEOGRAPHIC SCALES
P (GEOGRAPHY | RACE)
IN CH FI JA KO VI
SINGH
WANG
CRUZ
SATO
KIM
NGUYEN
Rows add up to 1
Columns add up to 1
TRACT 1
TRACT 2
TRACT 3
TRACT 4
TRACT 5
TRACT 6

Fig. 1 Flowchart summarizing the main data workflow to produce our published data, as detailed in the Methods section.

different probabilities, and primarily leverage this technique to update the prior probability of race given surname with the probability of location given race (one algorithm adds a second update on the probability of first name given race). Finally, the relationship between names and race is not one-to-one, as name lists suggest; nor is it permanent or fixed. For example, Chinese and Korean Americans share common surnames[57], which name lists can only assign wholesale to one group or the other. There is also substantial interracial marriage between Asians and other major race groups[58] and among Asian subgroups[59], both of which undermine the assumption that some first and last names can be exclusively associated with particular Asian subgroups.

Despite all of these reasons to prefer a name algorithm approach for Asian subgroup disparity assessments, the leading name algorithms cannot actually be used to predict Asian subgroup membership in their current form. In fact, because they primarily rely on the U.S. Census Bureau's 2010 Decennial Census surname data[21], which still uses 1977 federal standards, they can't even predict whether somebody is Asian – only the broader API. Alternative sources of name-race distribution information like mortgage application files from the Consumer Financial Protection Bureau[60] and voter files[61] also fail to report disaggregated data for Asian and NHPI groups consistently. These challenges highlight the need for name-race distribution data at the granularity of Asian subgroups.

Our contribution. Our work fills this critical data gap in racial equity research by providing the first dataset of name-race distributions for Asian subgroups, which can be directly used in combination with publicly available distributions of Asian subgroups by geographies to predict subgroup membership. Leveraging publicly available name and demographic information from Wikidata, we query data for 308,296 individuals from six Asian countries, extract frequencies of 25,876 first names and 18,703 surnames, and construct conditional probability tables that can be used as a proxy for U.S. name-race distributions among the six largest Asian subgroups: Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese. We validate the predictive performance of a disaggregated version of BISG using a rich electronic health records dataset with recorded subgroup-level race data for patients across the U.S. Despite the proxy nature of name distributions from an international and biased (*i.e.*, those notable enough to be in Wikidata) cohort, we find that this name-race distribution data outperforms deterministic name lists from SSA on most prediction tasks for our U.S. validation set, and can generate meaningful estimates of Asian health disparities. We hope these methods and findings, and our released dataset, demonstrate the promise of leveraging crowdsourced and continuously growing knowledge bases like Wikidata for practical improvements in racial disparity assessments.

Methods

In this section, we describe our methodology for querying Wikidata, extracting first and last name distributions, and producing probability tables, as illustrated in Fig. 1. We also detail our race imputation procedure, as implemented in the Technical Validation section, and outline alternative specifications users may consider to adapt this workflow for different purposes.

Wikidata. We use Wikidata[62] (https://query.wikidata.org), an open-source database of "items" with structured and queryable "properties" (see Table 1) – most importantly, family name and given name for non-fictional people – in Wikimedia, which includes Wikipedia. Ideally, we would like to query only the first names and surnames of U.S. residents who are members of our six Asian subgroups of interest, but we are severely limited by the availability of individuals who have the structured properties of both U.S. residency and Asian subgroup membership. In fact, the property of racial or ethnic group is sparsely populated in Wikidata (<1% in our study sample), given the consensus among the open-source community that a very high standard of proof is needed for this field to be used. As an alternative, we could consider an individual's country of birth and/or citizenship, which have been considered viable proxies for race in prior research[13], despite some obvious imperfections (*i.e.*, immigration between Asian countries[63]). We ultimately query people born between 1800 and 2020 with the properties of birthplace or citizenship in India, the People's Republic of China, the Philippines, Japan, South Korea, and Vietnam. Our queries for this study were run on September 23, 2024. An example query is provided below, specifically for all persons born 1950–1954 with birthplace or citizenship in Japan:

```
SELECT DISTINCT ?fullname ?fullnameLabel ?surnameLabel ?firstnameLabel
?dobLabel ?citizenshipLabel ?birthplaceLabel ?article
WHERE {
 ?fullname wdt:P31 wd:Q5 .
 ?fullname wdt:P569 ?dob .
 FILTER(YEAR(?dob) >= 1950 && YEAR(?dob) <= 1954)
 {?fullname wdt:P27|(wdt:P19/wdt:P17) wd:Q17 .}
 OPTIONAL {?fullname wdt:P734 ?surname}
 OPTIONAL {?fullname wdt:P735 ?firstname}
 OPTIONAL {?fullname wdt:P569 ?dob}
 OPTIONAL {?fullname wdt:P27 $citizenship}
 OPTIONAL {?fullname wdt:P19/wdt:P17 $birthplace}
 OPTIONAL { ?fullname wdt:P172 ?race }
 OPTIONAL {
  ?article schema:about ?fullname .
  ?article schema:inLanguage "en" .
  ?article schema:isPartOf <https://en.wikipedia.org/> .
 }
 SERVICE wikibase:label {bd:serviceParam wikibase:language
"[AUTO_LANGUAGE],en" .}
}
```

Table 2 shows sample outputs from our full Wikidata search query. The ID refers to the unique identifier of a person in Wikidata. The full name is the English label of a person's entry in Wikimedia, *i.e.*, the English title of their Wikipedia page if one exists, and is missing if a record has no English page. This full name label may exactly match the more structured surname and first name labels, but in many cases differs substantially, and all three

Description	Wikidata code
Non-fictional person	Q5
Instance of	P31
Date of birth	P569
Family name	P734
Given name	P735
Racial or ethnic group	P172
Country	P17
Birthplace	P19
Citizenship	P27
India	Q668
People's Republic of China	Q148
Phillipines	Q928
Japan	Q17
South Korea	Q884
Vietnam	Q881

Table 1. Wikidata codes for items and properties relevant to our study.

ID	Full Name	Surname	First Name	Subgroup
Q5625451	Leela Gandhi	Gandhi	Leela	Asian Indian
Q6408733	Kim Ho-Dong		Ho-Dong	Korean
Q12961160	Agueda Esteban	Esteban		Filipino
Q487161	Han Yongun	Han	Yongwoon	Korean
Q41152	Akebono Taro	Rowan	Chad	Japanese
Q9134318	Zhuang Qiaosheng			Chinese
Q110655352	Q110655352	Nguyen		Vietnamese

Table 2. Examples of individuals identified by Wikidata search query.

fields generally vary in completeness and quality, thus requiring further cleaning and processing. Overall, we identified over 300,000 unique individuals who matched our queries based on birthdate, birthplace, and citizenship, and who had string information in at least one of the three name fields.

Data cleaning and processing. We perform data cleaning and processing using R version 4.1.2. Following the general character string cleaning practice of the Census Bureau[21], we convert character encoding to ASCII Latin, remove punctuation and standalone characters, and cast all names to uppercase (similar cleaning should be applied to the names for which one is imputing race).

Next, we extract a single first and last name for each individual across multiple steps. First, any cleaned strings in the structured first name (45% complete) and surname (51% complete) fields are directly selected, with spaces removed. Second, we look to the cleaned string in the full name fields for candidates to fill missing first or last names. We consider single words, as well as contiguous, concatenated word pairs and triplets (*i.e.*, we are able to consider DELACRUZ if it were written as DE LA CRUZ), as candidates if they match names from a large existing name-race distribution dataset[61] (https://doi.org/10.7910/DVN/SGKW0K), which combines surnames from the Census Bureau 2010 surname list with first and last names from voter records, but only provides probabilities associated with the broad API category. We then rank multiple candidates based on the product of their $P(race|name)$ API score from that dataset (*i.e.*, prefer names more precisely associated with API), their $P(name|race)$ API score (*i.e.*, prefer names more prevalent among API), and character length of name (*i.e.*, prefer DELACRUZ over CRUZ, all else being equal), and select the highest scoring candidate. We perform this selection for a missing surname before selecting a missing first name; when selecting a first name candidate, we exclude candidates that match an already selected surname, and vice versa. Third, if either the first or last name is still missing, and the full name field yields a single candidate after the prior process of elimination, we select that remaining word. While prior work[61] filtered out names that appeared fewer than 25 times for privacy protection, we do not apply this same filter because of the public nature of Wikidata. Instead, we keep all first and last names that match Census Bureau and voter record names, and only filter out the remaining names if they appear only once, since they likely include a large proportion of practical errors. After these steps, our sample contains 308,296 individuals with 25,876 unique first names and 18,703 unique surnames.

Finally, in response to the practical imbalances in name frequencies across subgroups – stemming from both the unequal prevalence of Wikidata entries (*i.e.*, there happen to be more queried people from Japan than from all five other countries combined) and the difference in proportions of country populations relative to proportions of related American subgroups – we perform a simple calibration by re-weighting the total count of

	Asian Indian	Chinese	Filipino	Japanese	Korean	Vietnamese
1	Singh	Wang	Cruz	Sato	Kim	Nguyen
2	Kumar	Li	Santos	Suzuki	Lee	Tran
3	Rao	Zhang	Reyes	Tanaka	Park	Le
4	Sharma	Chen	Garcia	Ito	Choi	Pham
5	Khan	Liu	Lopez	Takahashi	Jeong	Hoang

Table 3. Top 5 most frequent surnames for each subgroup, as identified by Wikidata search query.

individuals per subgroup according to the counts of each Asian subgroup in the U.S. from the 2022 5-year American Community Survey (ACS), specifically Table B02015 (https://data.census.gov). From there, we produce probability tables of the format used in name algorithms, $P(race|name)$ and $P(name|race)$, by performing row-wise and column-wise normalization on our name-subgroup pairs, respectively. This is the output that we publish, as described in the Data Records section. Table 3 lists the top five most frequent surnames for each subgroup.

Application of BISG. To demonstrate application of our published dataset, and as part of our technical validation, we implement conventional BISG[3,4], which generates a posterior probability that an individual i is identified with a race R_j, given their surname S_i and geographic area G_i, using Eq. 1 where m is the number of racial categories:

$$P(R_j|S_i,\ G_i) = \frac{P(R_j|S_i)\cdot P(G_i|R_j)}{\sum_{j=1}^{m} P(R_j|S_i)\cdot P(G_i|R_j)} \tag{1}$$

$P(R_j|S_i)$ denotes the conditional probability that individual i is identified with race group R_j given their surname S_i, as retrieved from our Wikidata name-race tables. $P(G_i|R_j)$ denotes the conditional probability that individual i lives in location G_i given their race R_j. G_i is typically chosen to be a census geography, in which case an individual's address is likely geocoded to a chosen census geography level, and the conditional probabilities are derived from ACS data. While the ACS Table B03002 is typically used to obtain geography-race frequencies based on federal race categories, we use Table B02015 which provides frequencies for different Asian subgroups.

We illustrate the performance of BISG's individual-level race predictions using precision-recall (PR) curves, which represent precision on the y-axis (the proportion of true positive cases out of all predicted positive cases) and recall on the x-axis (the proportion of predicted true positive cases out of all true positive cases) at different classification thresholds (the posterior probabilities from the name algorithm), where each prediction is made for one subgroup versus the rest. We bootstrap 100 times, resampling all patients with replacement, and plot the mean precision (along with the 2.5% and 97.5% percentile precision as a confidence interval) across recall values.

To produce a subgroup-level probabilistic estimate of prevalence[12,54,64], we use Eq. 2, which calculates the prevalence of outcome Y (1 denotes the presence of outcome, and 0 otherwise) for individuals in race group j by weighting each individual's contribution by the posterior probability p_{ij} that individual i belongs to subgroup j.

$$\widehat{prevalence}_j^{Y} = \frac{\sum_i^N p_{ij}\cdot Y_i}{\sum_i^N p_{ij}} \tag{2}$$

Alternative specifications. We explored many alternative choices in the pipeline, from Wikidata querying, to name processing, to name algorithm implementation, to the formulation of the individual-level race predictions and group-level disparity predictions, and finally to alternative name data sources. In the Technical Validation section, we compare a range of these alternatives to the main results. While they generally did not perform as well in our setting, they may offer nuanced improvements for users in different settings. If a large enough validation set is available, we recommend constructing a "dev-test" split, in which the various alternative approaches outlined below are evaluated using the dev set, and then the best approaches are confirmed to perform well on the test set before being utilized on individuals with missing subgroup information.

Querying. Wikidata enables a fundamentally open-ended degree of querying, enabling researchers to add considerably more specificity to their searches than we describe here. For example, if the individuals for whom race is being predicted are known to be of a certain gender, or of a certain age range, a filter on structured gender information or year of birth could be added. However, filters for specificity may trade off with sufficient availability of names. It is also possible to query all direct descendants of a particular individual. Descendants may have different countries of birth or citizenship but names that can be associated with the same racial subgroup as an ancestor. However, as a result of interracial marriage, descendants may also identify with different racial subgroups as their ancestors.

We also emphasize that the specific set of countries (as locations of birthplace or citizenship) considered as proxies for race can be modified to more precisely account for geopolitical changes that occurred across these regions between 1800 and 2020[65–70]. For example, while we query only for modern-day South Korea in our main approach as a proxy for the Korean American subgroup, yielding about 38,000 Wikidata individuals, adding a query for modern-day North Korea (*i.e.*, Democratic People's Republic of Korea) yields 2,400 individuals, 1,800 of whom are newly identified. Furthermore, adding a query for the pre-20th century Korean dynasty yields another 1,600 individuals, 1,100 of whom are newly identified. Including these other queries may or may not improve predictive performance, depending on the specific individuals for whom Korean American subgroup identification is being predicted. For our published dataset, we make the conservative choice of only querying a single contemporary Asian nation for each subgroup, but users are encouraged to adjust these choices to better reflect the nuanced historical complexities[71] that may shape their particular research question.

Name processing. Our main approach involves a multi-step process of extracting a first and last name for each individual, prioritizing any existing information in the structured first name and surname fields before then considering and selecting from candidates in the unstructured full name field. A possible "minimal" approach instead only relies on the structured first name and surname fields, thereby concluding with fewer distinct names. A possible "maximal" approach considers candidates from the unstructured full name field, but then does not use a ranking procedure to select the best candidate; instead, all candidates are preserved and given equal fractional weight at the person-level in the subsequent step of producing probability tables. In other words, while the number of Wikidata individuals does not change, the total number of distinct names – many of which are clearly lower quality because they are concatenated word pairs or triplets from the full name field – significantly increases. Recall that in any of these approaches, distinct names with only a single observation from Wikidata are excluded, and the final resulting probability tables are normalized such that $P(race/name)$ and $P(name/race)$ add up to 1 row-wise and column-wise, respectively.

Name algorithm implementation. Given completed probability tables, users may consider different choices in the specific name algorithm used to produce posterior probabilities for individuals.

First, users can consider utilizing different scales of geography in BISG. Prior work[72] suggests that more granular geographic information improves the performance of race imputation, but ACS data, due to privacy protection and sampling uncertainty, often incorrectly reports zero counts for minority groups (*e.g.*, Asians)[73]. This zero-count problem becomes even more pronounced when computing the racial composition of smaller minority groups, such as Japanese and Vietnamese, at the census tract level, the most granular geographic unit at which ACS provides population counts for Asian subgroups[74]. While we use the county version of geography-race distribution tables in the Technical Validation section, such distributions are also available at other census geographies, the most granular option being the census tract level for Table B02015.

Second, users can consider incorporating first name information. Bayesian Improved First Name Surname Geocoding (BIFSG)[75] generates a posterior probability that an individual i is identified with a race R_j, given their first name F_i, surname S_i, and geographic area G_i with Eq. 3, where m is the number of racial categories:

$$P(R_j|F_i,\,S_i,\,G_i)=\frac{P(R_j|S_i)\,\cdot\,P(G_i|R_j)\,\cdot\,P(F_i|R_j)}{\sum_{j=1}^{m}(R_j|S_i)\,\cdot\,P(G_i|R_j)\,\cdot\,P(F_i|R_j)} \tag{3}$$

$P(R_j|S_i)$ denotes the conditional probability that individual i is identified with race group R_j given their surname S_i. $P(F_i|R_j)$ denotes the conditional probability that individual i has first name F_i given their race R_j. These two quantities are derived from our name-race tables.

Subgroup-level health disparity prediction. Prior work[52] suggests that the weighted estimator in Eq. 2 reduces bias in disparity analysis, compared to estimators that use categorical classification. However, the main identifying assumption of this estimator, namely that conditional on predicted race, there is no residual correlation between the outcome and actual race, may be violated in many settings, introducing bias in our estimates.

Users might explore a variety of alternative approaches to estimating subgroup-level outcomes. For instance, posterior probabilities can be converted into classifications, the most straightforward approach being to assign a single subgroup identification for each individual based on the highest posterior probability. This approach and the weighted estimator in Eq. 2 share the characteristic of utilizing all individuals in the cohort to estimate subgroup-level outcomes.

Alternatively, a subset of individuals can be utilized for the estimates, in a variety of ways. First, a classification threshold (*e.g.*, 90%) can be set, restricting classification to only individuals with a posterior probability above that threshold within each subgroup. Second, a count threshold (*e.g.*, 1000 individuals) can be set, restricting selection to only that number of individuals within each subgroup, in descending order of posterior probability. Both of these approaches have the effect of favoring individuals with the highest probability of being correctly predicted, though as seen in the precision-recall curves, the highest average precision may not necessarily be achieved at extremely high classification thresholds. Either of these approaches may also be implemented with either the weighted estimator in Eq. 2 or single-subgroup classifications. It's worth noting that, depending on the thresholds set, individuals may contribute to more than one subgroup's group-level estimates.

Historical census names. One possible alternative to using Wikidata names is using names from historical U.S. census surveys. However, due to the 72-year moratorium on releasing confidential information about individuals from census records, first and last names are only available from 1950 and prior. Nonetheless, we collected publicly accessible census samples from IPUMS[76] (https://usa.ipums.org), filtered to only individuals who identified as one of the six major Asian subgroups or provided their birthplace as one of the six Asian countries, and cleaned and processed their names using a similar probabilistic approach as our main approach. Our final alternative dataset represents (at most) 161,485 individuals and contains 10,758 unique first names and 10,431 unique surnames, which can be converted into name-race tables following the aforementioned approaches.

Data Records

The data records are available for download at Harvard Dataverse[77]. All name-race tables have seven columns. The first column in each dataset represents the uppercase first or last name. The other six columns provide the conditional probabilities for Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese subgroup membership, as derived from Wikidata. Although the necessary geography data for name algorithm implementation is publicly available, we include for convenience the properly formatted geography-race distribution tables, $P(race|geography)$ and $P(geography|race)$ for different geographical levels. These geography-race tables are similarly structured, with the main difference being that the first column represents the geographic identifier (GEOID). All tables are provided in R data and CSV formats.

We also provide raw data files and R scripts in our GitHub repository (https://github.com/reglab/disaggregation), and invite users to explore alternative approaches, as detailed in the Methods section.

Technical Validation

In this section, we compare the performance of our name-race data with the SSA name lists used in previous studies and demonstrate a use case of our data in a healthcare research setting.

Comparison with SSA name lists. The SSA name lists[13] draw information from 1.6 million Asian Americans from the six main subgroups who were born in Asia before 1941, and contain 11,291 unique first names and 20,693 unique surnames. 5,758 first names and 7,333 surnames appear in both the SSA name lists and our Wikidata name tables. The SSA name lists retain 5,533 unique first names and 13,360 unique surnames uncaptured by our Wikidata name tables, while our Wikidata name tables conversely add 20,118 new first names and 11,370 new surnames.

To assess the relative performance of these two name sources, we can control for the effect of additional geography information by incorporating SSA surnames as an input into the same BISG name algorithm. In effect, if an individual has a name that matches the SSA surname list, their posterior probability will be 100% for the matching race, and 0% for all other races – any available geography information will not affect this. If an individual has a name that does not match the SSA surname list, their posterior probability will be the same as the geography prior. Comparing SSA and Wikidata approaches, individuals with names that match to neither source will have exactly the same prediction based on geography. Therefore, any overall performance difference will come down to relative differences in three subsets of the evaluated population: individuals with names that match to *(1)* both SSA and Wikidata, *(2)* only SSA, and *(3)* only Wikidata.

For subset *(1)*, comparing the names in common, if we classify the Wikidata names based on the race category with the highest posterior probability, we achieve 95% agreement on first names and 93% agreement on surnames. In effect, for these names in common, we can successfully recover deterministic information using Wikidata; on the other hand, we cannot successfully recover probabilistic information using the deterministic SSA name lists. As for the probabilistic information gained, 15% of the shared first names and 14% of the shared surnames were observed across multiple subgroups in Wikidata, allowing them to be represented probabilistically in the name-race distribution tables. Performance differences in this subset will thus largely hinge on the impact of probabilistic information for these particular names, which may themselves be more common in present-day populations.

Given that subsets *(2)* and *(3)* may yield mutually exclusive performance benefits, users might very well consider combining the two sources, or combining name information from multiple sources generally. While merging multiple sources creates a considerable challenge for the estimation of $P(name|race)$ because different sources are not necessarily bound to the same base rate of coverage, our BISG implementation does not require this permutation of the surname probability tables at all, and instead, $P(race|name)$ can easily incorporate multiple mutually exclusive sets of names with either deterministic race labels or probabilistic distributions. We ultimately evaluate a hybrid approach that matches the Wikidata approach for subset *(1)* and the SSA approach for all other subsets.

Electronic health records. We demonstrate the applicability of our new dataset in a healthcare research setting. The American Family Cohort is a collection of electronic health records (EHRs) of over eight million patients from 2010 to 2024[78]. This dataset is well-suited for our validation task because it contains first name, surname, and residential information to construct BI(F)SG estimates, and recorded race at the Asian subgroup level to assess the performance of imputations. Compared with other EHR data sources, AFC over-samples historically underserved populations including rural, low-income, and minority groups. This research was approved by the Institutional Review Board, and all data analyses were performed on servers with approvals for high-risk data and protected health information (PHI).

Over 170,000 patients in the EHR data are recorded as Asian, but only 22,182 Asian patients are recorded with subgroup membership in at least one of the six major subgroups, which illustrates the dearth of disaggregated race reporting our dataset seeks to address. For the purposes of the most foundational technical validation, we select only the 19,921 patients who are recorded as members of exactly one of the six subgroups, and who also have a recorded surname, as our validation set. The relative makeup of these patients (7% Asian Indian, 18%

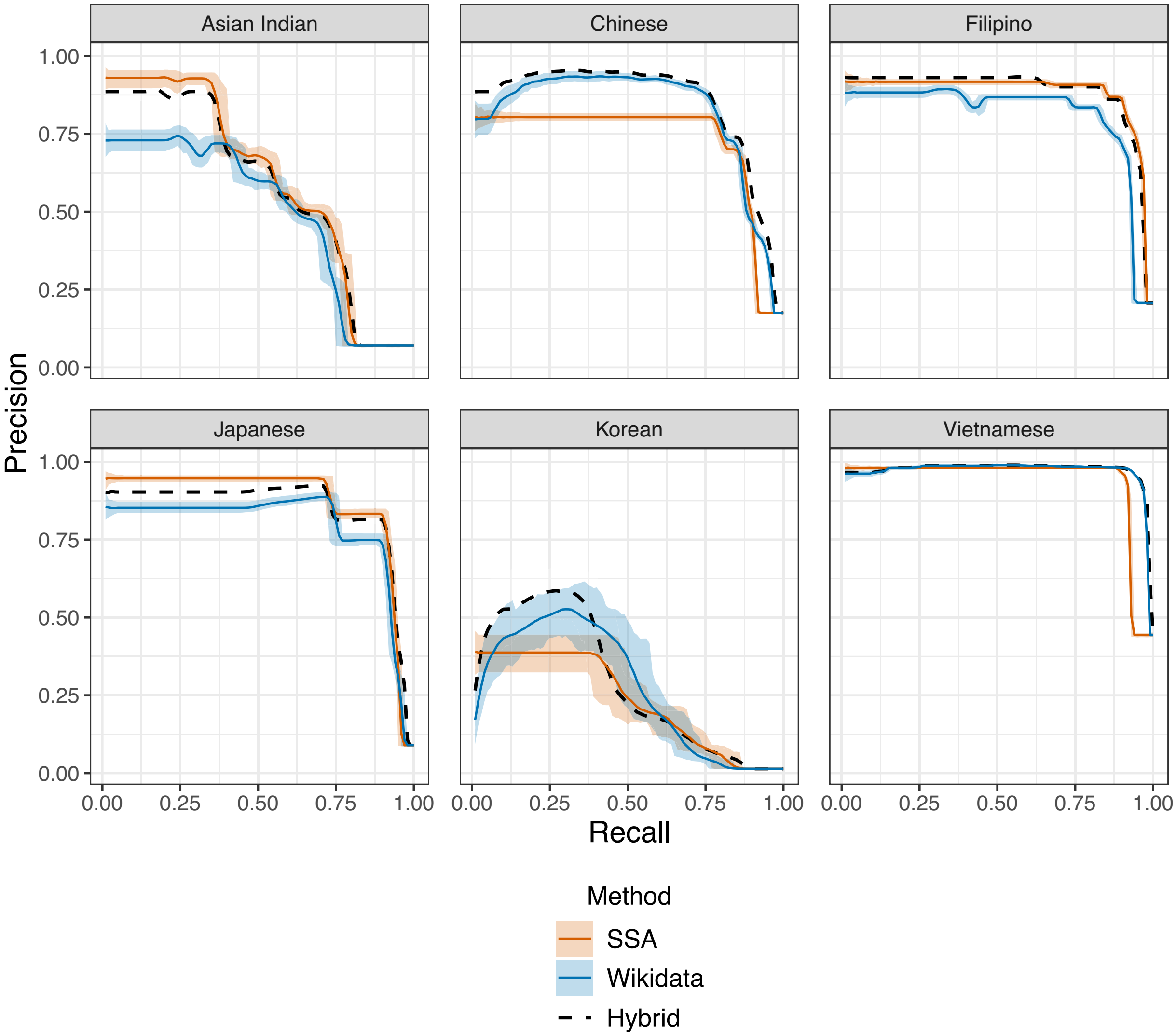

Fig. 2 Precision-Recall Curves, comparing deterministic name information sourced from SSA (orange), probabilistic name information sourced from Wikidata (blue), and a hybrid approach using both sources (dashed). The shaded region denotes the 95% confidence interval based on 100 bootstrapped PR curves. Intervals for the hybrid approach are omitted for legibility.

Chinese, 21% Filipino, 9% Japanese, 1% Korean, and 44% Vietnamese) differs from overall U.S. proportions (27% Asian Indian, 28% Chinese, 19% Filipino, 5% Japanese, 9% Korean, and 12% Vietnamese). However, it nonetheless reflects a real-world sample which may realistically be used for health disparity assessments.

Individual-level race prediction. Figure 2 compares the precision-recall (PR) curves of predictions, by subgroup, using the two surname sources as inputs to BISG. We observe that the SSA PR curves each feature an initial plateau, the rightmost edge of which essentially represents the single pair of precision and recall that would be achieved if one were to only select deterministically matched SSA names. This point represents the highest possible classification threshold (100%), indiscriminately applied to all deterministic matches, and thus its coordinate reflects the single pairing of precision and recall performance expected if the deterministic information from SSA were conventionally used as in prior literature, and the plateau itself represents a practical ceiling on possible precision using SSA. The Wikidata PR curves for Chinese and Korean predictions achieve significantly higher levels of precision within a certain range of recall, while the hybrid approach roughly matches the better of the two approaches for each subgroup.

The PR curves for Korean predictions particularly stand out for their wide confidence intervals, consistent with the fact that Koreans are the most underrepresented subgroup in our validation set. Korean predictions also yield a lower average precision across recall levels, which is likely driven by three distinct factors, among others. First, the $P(race|surname)$ term in BISG systematically favors a Chinese prediction for some surnames (*e.g.*, Han, Song, Yang) that are common among both Chinese and Korean individuals[57] on account of the population-level ratio of Chinese to Korean Americans (approximately 3:1), which degrades the performance of Korean predictions for those surnames. Second, our small sample of Koreans in the validation set may also

Subset		Asian Indian	Chinese	Filipino	Japanese	Korean	Vietnamese
Match to both	*n*	416	2,993	1,717	1,338	168	8,445
	SSA	0.57 (0.52–0.62)	0.73 (0.72–0.75)	0.84 (0.82–0.85)	0.89 (0.88–0.91)	0.29 (0.24–0.33)	0.94 (0.93–0.94)
	Wikidata	0.56 (0.51–0.62)	0.89 (0.87–0.90)	0.84 (0.82–0.86)	0.87 (0.85–0.89)	0.41 (0.34–0.48)	0.97 (0.96–0.97)
Match to Wikidata only	*n*	319	210	310	120	58	201
	SSA	0.47 (0.40–0.52)	0.26 (0.22–0.30)	0.67 (0.61–0.72)	0.43 (0.35–0.51)	0.10 (0.06–0.13)	0.54 (0.49–0.60)
	Wikidata	0.42 (0.37–0.48)	0.33 (0.28–0.38)	0.36 (0.32–0.41)	0.35 (0.27–0.41)	0.17 (0.11–0.24)	0.59 (0.53–0.67)
Match to SSA only	*n*	29	64	1,095	42	1	53
	SSA	0.31 (0.15–0.47)	0.72 (0.61–0.82)	0.95 (0.94–0.96)	0.72 (0.63–0.87)		0.73 (0.61–0.86)
	Wikidata	0.18 (0.09–0.29)	0.16 (0.12–0.22)	0.97 (0.96–0.97)	0.15 (0.11–0.19)		0.50 (0.38–0.62)
Match to neither	*n*	635	235	1,001	289	56	126
	SSA	0.62 (0.59–0.65)	0.19 (0.17–0.23)	0.77 (0.75–0.80)	0.45 (0.40–0.49)	0.04 (0.03–0.05)	0.35 (0.29–0.41)
	Wikidata	0.62 (0.59–0.65)	0.19 (0.17–0.23)	0.77 (0.75–0.80)	0.45 (0.40–0.49)	0.04 (0.03–0.05)	0.35 (0.29–0.41)
All patients	*n*	1,339	3,502	4,123	1,789	283	8,825
	SSA	0.59 (0.56–0.61)	0.72 (0.70–0.73)	0.87 (0.86–0.88)	0.86 (0.84–0.87)	0.23 (0.18–0.27)	0.93 (0.92–0.93)
	Wikidata	0.49 (0.47–0.51)	0.81 (0.79–0.82)	0.80 (0.79–0.81)	0.78 (0.76–0.80)	0.27 (0.22–0.32)	0.96 (0.95–0.96)
	Hybrid	0.57 (0.55–0.60)	0.84 (0.83–0.85)	0.87 (0.86–0.88)	0.83 (0.82–0.85)	0.28 (0.23–0.33)	0.96 (0.96–0.96)

Table 4. Average precision by subgroup for all 19,921 patients in the validation set as well as four mutually exclusive subsets: 15,074 patients with names that match to both SSA and Wikidata sources; 1,215 patients with names that match to Wikidata only (*i.e.*, the SSA values provided are only geography-based predictions); 1,287 patients with names that match to SSA only (*i.e.*, the Wikidata values provided are only geography-based predictions); and 3,550 patients with names that match to neither source (*i.e.*, both values are only geography-based predictions and therefore match exactly). The hybrid approach is the Wikidata approach for names that match to both SSA and Wikidata, combined with the SSA approach for all other cases. Provided 95% confidence intervals are derived from 100 bootstraps. Some performance results are omitted due to sample size.

happen to feature a less representative distribution of names, which our bootstrapping procedure cannot adequately account for. Third, the names we extract from Wikidata and assign to the Korean subgroup may also be less representative of real Korean American names and name frequencies than we find for names assigned to other subgroups.

Table 4 provides a closer look at the relative performance of SSA versus Wikidata for key patient subsets of overlapping and non-overlapping name coverage, which each contribute to overall relative performance. The evaluative metric, average precision, is conceptually the area under the PR curves as shown in Fig. 2, and is more specifically a weighted average of the precision values at each threshold, where the weights are the increase in recall from the previous threshold. For the 15,074 (71% of) patients with names that match to both SSA and Wikidata sources, the probabilistic information in Wikidata performs on par with the deterministic information in SSA across Asian Indian, Filipino, and Japanese predictions, but otherwise outperforms the deterministic information in SSA across Chinese, Korean, and Vietnamese predictions, with differences being statistically significant. Results for this subset validate that the probabilistic information on names obtained from Wikidata provide predictive value, though performance varies substantially by subgroup. As noted previously, with our current query approach, for all the 7,333 unique surnames that match to both SSA and Wikidata, 14% gain probabilistic information as a result of the Wikidata approach. However, in this subset of 15,074 patients from our validation set, the actual sample of surnames is far from a uniform distribution across the 7,333; instead, 77% (11,670) of these patients have surnames with probabilistic information from Wikidata. This suggests that while Wikidata only provides enriched information for a minority of the unique surnames it shares with SSA, those surnames are substantially more common in the present-day population, and enrichment with probabilistic information, *i.e.*, replacing the deterministic prediction SSA would have made, results in clear performance gains for three subgroups.

For the additional 1,215 (6% of) patients with names captured by Wikidata, but not by SSA, the comparison is effectively between BISG with and without the surname prior. Surprisingly, the probabilistic information in Wikidata is far less precise for this subset than it was in the first subset, and for Asian Indian, Filipino, and Japanese predictions, performance with just geography priors is in fact substantially better than performance with surname priors. In other words, for patients with these names, one is better off not taking into account name information at all. Therefore, our hybrid approach does not make use of this subset of names from Wikidata, though users may find these names useful in other settings.

For the next 1,287 (6% of) patients, for whom names match to SSA but not Wikidata, the SSA approach does maintain strong predictive performance relative to just using geography priors (excluding Korean, for which insufficient data is available). Interestingly, geography-based prediction once again outperforms name-based prediction for Filipino, and, at 97%, exhibits the highest precision we observe in the analysis (tied with Wikidata-based prediction for Vietnamese in the first subset). Indeed, 86% of the Filipinos in this subset live in counties where Filipinos outnumber the other five Asian subgroups, so performance here is likely reflective of relatively monocultural communities.

The remaining 3,550 (17% of) patients match to neither name source and so, as previously noted, exhibit no difference in predictive performance. Filipino predictions are once again relatively precise for being solely geography-based. In summary, the Wikidata approach outperforms SSA for Chinese, Korean, and Vietnamese predictions but underperforms SSA for Asian Indian, Filipino, and Japanese predictions. The hybrid approach recovers most of the losses relative to SSA while expanding the gains of Wikidata, including a statistically significant boost on Chinese prediction performance compared to either approach alone. These improvements can be attributed almost entirely to the 1,027 unique surnames Wikidata shares with SSA and enriches with probabilistic information, and which turn out to be the surnames of a majority of patients in our validation set.

As a last note about individual-level predictions, we emphasize that empowering researchers and practitioners to control the precision-recall tradeoff is substantively valuable. Consider a public health intervention, where a scarce resource is to be targeted to at most 100 individuals from a particular subgroup in our validation set. In this setting, precision should be prioritized, and the setting resembles interventions by public health officials during the Covid-19 pandemic to allocate scarce translation resources for contact tracing[79,80]. As seen in Fig. 2, depending on the subgroup, either the SSA, Wikidata, or hybrid approach may achieve a significantly higher maximum precision than the others. For example, by sampling 100 Chinese individuals with a hybrid-based posterior probability of 97% or higher, relative to sampling 100 Chinese individuals who match to the SSA name list, the number of Chinese individuals in that sample would increase by 15 [5–25, 95% confidence interval] to an average of 94. Conversely, some interventions prioritize recall, such as a public health outreach campaign aiming to reach nearly all members of a subgroup (*e.g.*[81]). For Vietnamese predictions, the Wikidata and hybrid approaches achieve near-perfect recall at a particularly high overall precision. Reaching 8,736 (99% of) Vietnamese individuals in the sample would require contacting 5,560 [4,150–6,620] fewer individuals using a hybrid-based posterior probability of 5% or higher, compared to the SSA approach. In short, there are strong reasons to prefer an approach that allows control over how to prioritize precision versus recall in research and real-world applications.

Subgroup-level health disparity prediction. We now return to the challenge that animates our work: assessing health disparities among Asian subgroups. The National Institutes of Health recently underscored the lack of disaggregated condition prevalence data for Asian subgroups across chronic health conditions, such as type 2 diabetes mellitus and hypertension[48]. We select these two illustrative conditions, as well as asthma and depression, which are emphasized by the Office of Minority Health within the U.S. Department of Health and Human Services[82], to evaluate whether our dataset can recover the true characteristics of disparities across a diversity of health conditions in our validation set of Asian American patients. The EHR data includes a history of clinical diagnoses for each patient, allowing us to code each of our 19,921 patients as having ever had each diagnosis, as mapped to a set of related SNOMED-CT codes[83]. We also include two relevant upstream testing rates available in the EHR data: whether each patient has ever had a hemoglobin A1c (HbA1c) test, one possible way to diagnose diabetes[84], and whether each patient has ever had a screening for depression[85].

For each outcome, using Eq. 2, we can produce six group prevalence estimates for the six Asian subgroups, as well as fifteen *predicted* pairwise disparities between the six subgroups. Meanwhile, we can directly calculate fifteen *actual* pairwise disparities using the ground truth subgroup labels in our validation set. Following prior work, we focus on the ability of our predictive methods to recover the maximum actual pairwise disparity between any two subgroups[86], as well as the average actual pairwise disparity across all subgroups.

Fig. 3 displays, for each of the six outcomes, subgroup-level prevalences, in order from highest to lowest as calculated using the actual EHR data (green), alongside the prevalence predictions using SSA (orange), Wikidata (blue), and a hybrid approach combining the two sources (black). Error bars are derived from our 100 bootstrap iterations. Table 5 details the aggregate prevalences calculated using the actual EHR data (from a low of 3.8% for depression screenings to a high of 24.7% for hypertension diagnoses), followed by a comparison between observed and predicted maximum and average pairwise disparities.

We highlight key takeaways from this validation of subgroup-level health disparity prediction. First, we observe substantial actual disparities across the six subgroups that would have been masked by the aggregated Asian category. For example, while the aggregate type 2 diabetes diagnosis prevalence for all Asian patients is 12.0%, the maximum subgroup disparity is almost as large a quantity, 10.4%, between Chinese patients (5.7%) and Japanese patients (15.7%, nearly triple the rate). We also note that Chinese patients have the lowest prevalence across five out of six outcomes, a trend consistently recovered by our predictions. For the remaining outcome, the HbA1c testing rate, Japanese patients exhibit the lowest prevalence; in other words, within these Asian subgroups, Japanese patients have both the highest risk and least likelihood of routine preventive checks for type 2 diabetes. While these particular observations captured in our EHR dataset may not necessarily reflect population-level health outcomes, other studies which have investigated a subset of these disparities find generally consistent results[48,82,87]. Other results presented here are less covered in existing literature, exemplifying the opportunity for new health research insights to be surfaced through the support of our methods.

Second, both predictive approaches generally capture the spread and rank ordering, and stay within the confidence bounds, of actual subgroup prevalences, though each approach makes various errors that suggest a high dimensionality of relationships across names, locations, and health outcomes. One consistent form of bias across the outcomes is an underestimate of the maximum subgroup disparity, *i.e.*, a regression towards the aggregate Asian prevalence, which may result from violations of the aforementioned identifying assumption of our probabilistic estimator.

Third, while prediction performance was more distinct among the SSA, Wikidata, and hybrid approaches in individual-level prediction tasks, for these group-level disparity predictions, performance differences are neither large nor statistically significant. More importantly, in the face of real disparities averaging as high as 15 percentage points among these six Asian subgroups, both name sources enable predictions that consistently capture the direction and magnitude of these disparities.

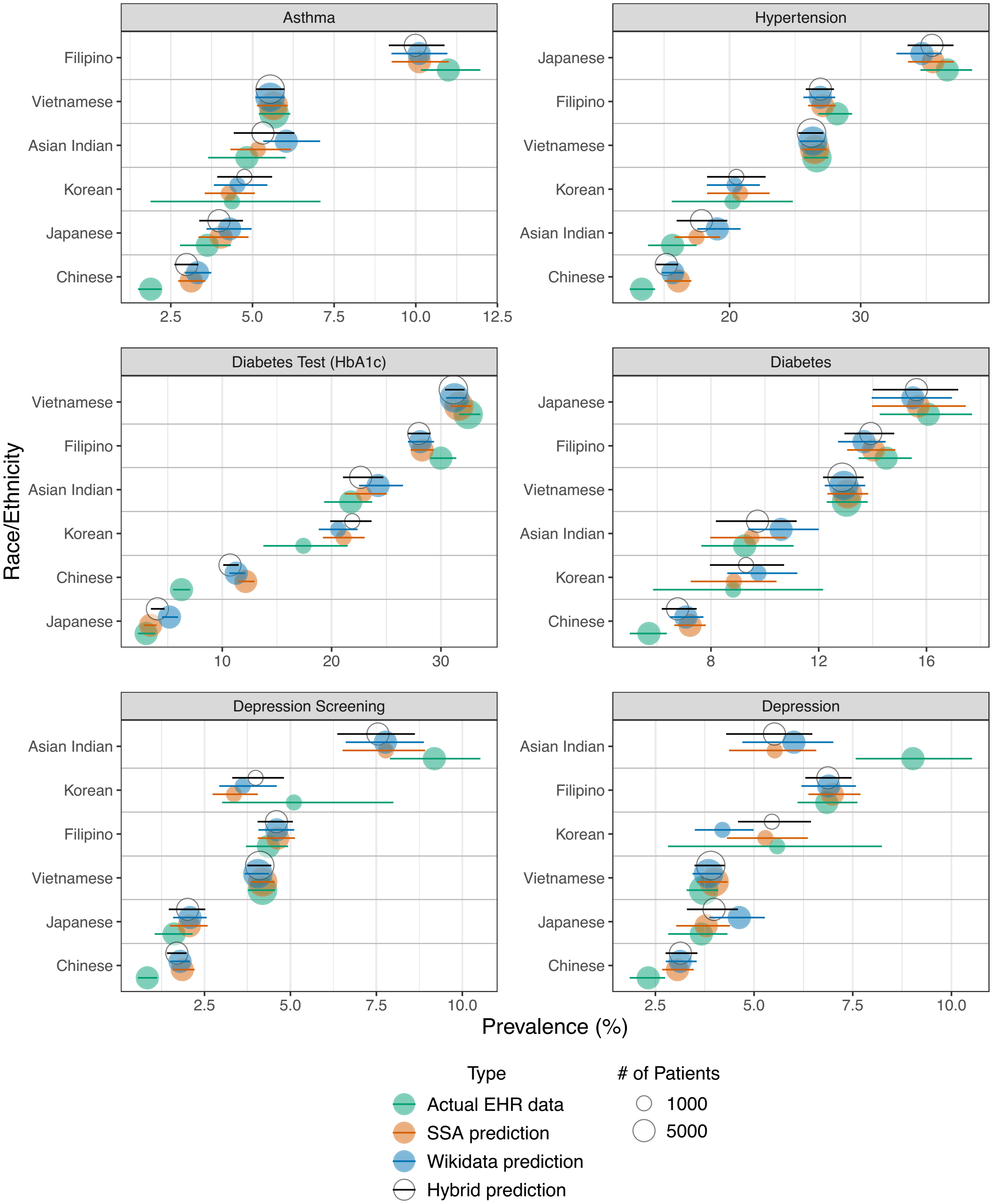

Fig. 3 Prevalences of six health diagnosis and testing outcomes among six Asian subgroups, as observed in EHR data (green), and as predicted using SSA (orange), Wikidata (blue), and a hybrid approach using both sources (black). The point sizes indicate subgroup size as observed or predicted, and the error bars denote the 95% confidence interval based on 100 bootstraps. Each plot is arranged in order of the observed prevalence for that condition.

Ultimately, our dataset allows researchers, policymakers, and other practitioners to reliably surface evidence of substantial disparities among six Asian subgroups, as well as leverage probabilistic information to improve the effectiveness of numerous equity-driven interventions.

		Asthma	Hypertension	Diabetes Test (HbA1c)	Diabetes	Depression Screening	Depression
Aggregate prevalence (%)	Observed	5.8 (5.6–6.2)	24.7 (24.0–25.3)	23.8 (23.4–24.3)	12.0 (11.6–12.4)	3.8 (3.5–4.0)	4.5 (4.3–4.8)
Maximum disparity between subgroups (%)	Observed	9.1 (8.2–10.1)	23.3 (20.9–25.5)	29.4 (28.0–30.7)	10.4 (8.6–12.2)	8.4 (7.1–9.6)	6.7 (5.2–8.2)
	SSA	7.0 (6.1–8.0)	19.5 (17.5–21.3)	28.2 (27.2–29.3)	8.5 (7.1–10.0)	6.0 (4.8–7.1)	3.9 (3.1–4.7)
	Wikidata	6.8 (5.8–7.7)	19.0 (17.0–20.8)	26.1 (24.7–27.4)	8.4 (6.9–9.9)	6.0 (4.9–7.2)	3.8 (3.0–4.5)
	Hybrid	7.0 (6.1–7.9)	20.2 (18.4–22.0)	27.1 (26.0–28.4)	8.9 (7.2–10.4)	5.9 (4.8–6.8)	3.7 (3.1–4.4)
Average disparity between subgroups (%)	Observed	3.6 (3.2–4.0)	10.7 (9.8–11.6)	14.9 (14.3–15.5)	4.9 (4.3–5.8)	3.6 (3.0–4.3)	3.1 (2.4–3.6)
	SSA	2.7 (2.4–3.1)	8.8 (8.0–9.6)	12.8 (12.3–13.2)	4.1 (3.4–4.8)	2.6 (2.1–3.0)	1.8 (1.4–2.2)
	Wikidata	2.7 (2.4–3.1)	8.3 (7.5–9.1)	12.3 (11.9–12.8)	3.8 (3.2–4.3)	2.5 (2.1–3.0)	1.7 (1.4–2.1)
	Hybrid	2.7 (2.4–3.0)	9.0 (8.1–9.7)	12.6 (12.1–13.0)	4.1 (3.5–4.7)	2.5 (2.0–2.9)	1.7 (1.4–2.1)

Table 5. Prevalence of six health diagnosis and testing outcomes, among 19,921 Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese patients. Observed results directly utilize information from the EHRs. The SSA and Wikidata approaches predict Asian subgroup using names from either source, respectively. The hybrid approach is the Wikidata approach for names that match to both SSA and Wikidata, combined with the SSA approach for all other cases. Provided 95% confidence intervals are derived from 100 bootstraps.

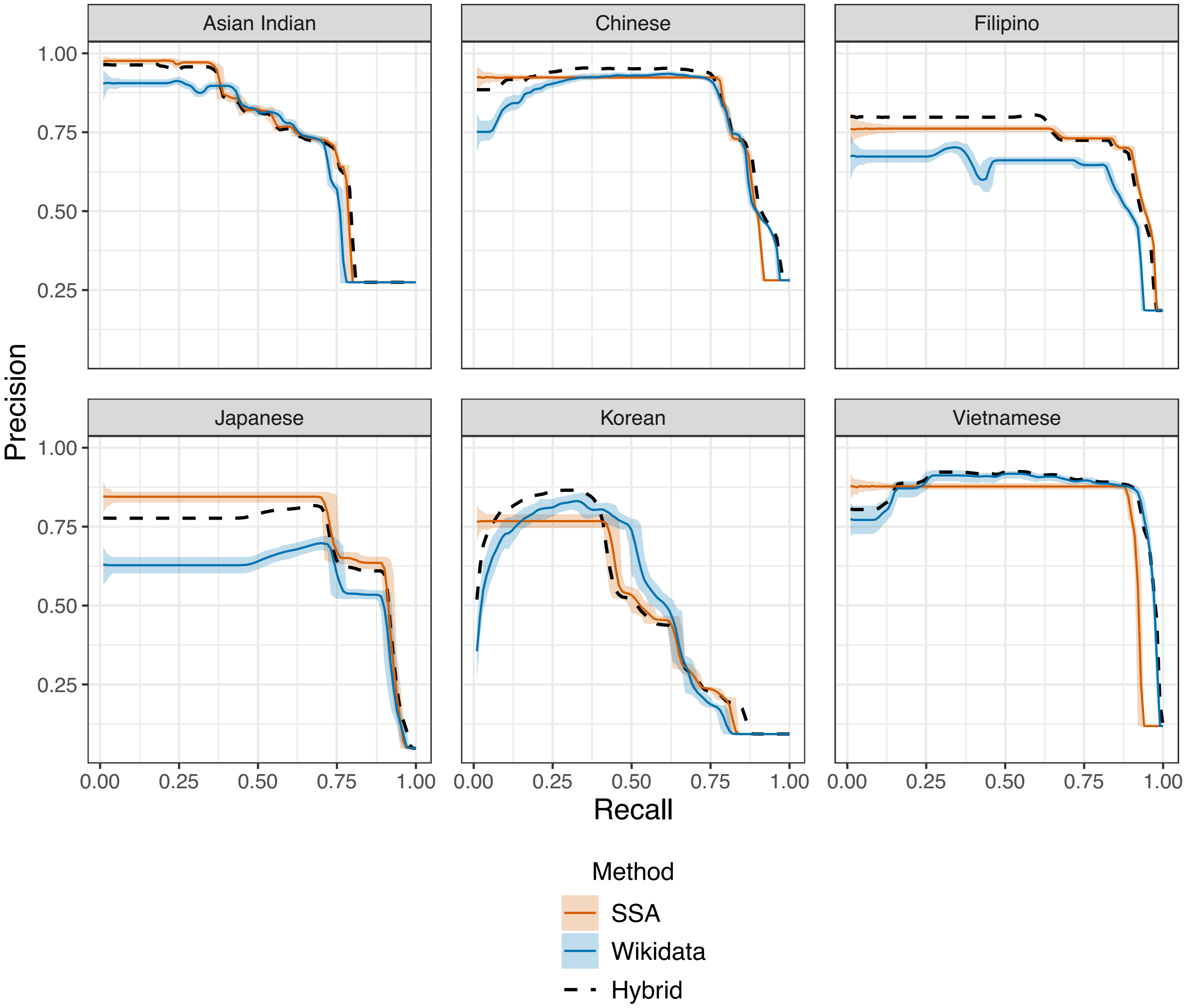

Fig. 4 Precision-Recall Curves, comparing deterministic name information sourced from SSA (orange), probabilistic name information sourced from Wikidata (blue), and a hybrid approach using both sources (dashed). Performance is evaluated on a version of the validation set with racial makeup balanced to U.S. proportions. The shaded region denotes the 95% confidence interval based on 100 bootstrapped PR curves. Intervals for the hybrid approach are omitted for legibility.

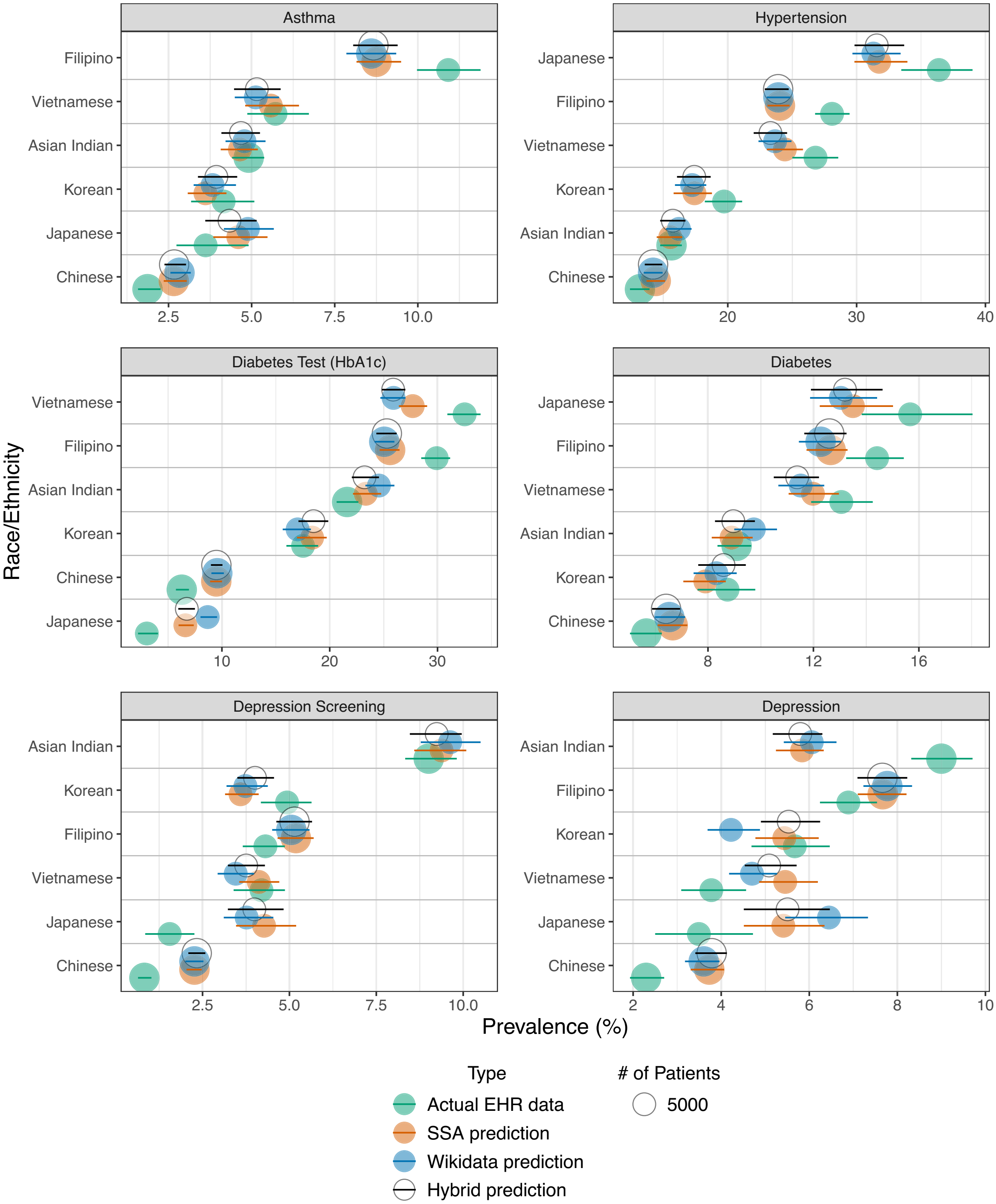

Fig. 5 Prevalences of six health diagnosis and testing outcomes among six Asian subgroups, as observed in EHR data (green), and as predicted using SSA (orange), Wikidata (blue), and a hybrid approach using both sources (black). Performance is evaluated on a version of the validation set with racial makeup balanced to U.S. proportions. The point sizes indicate subgroup size as observed or predicted, and the error bars denote the 95% confidence interval based on 100 bootstraps. Each plot is arranged in order of the observed prevalence for that condition.

Results using a balanced validation set. As previously noted, the relative makeup of the patients in our EHR validation set differs substantially from overall U.S. proportions of these Asian subgroups. We retain the existing makeup for the main technical validation because it reflects a real-world prediction setting, in which

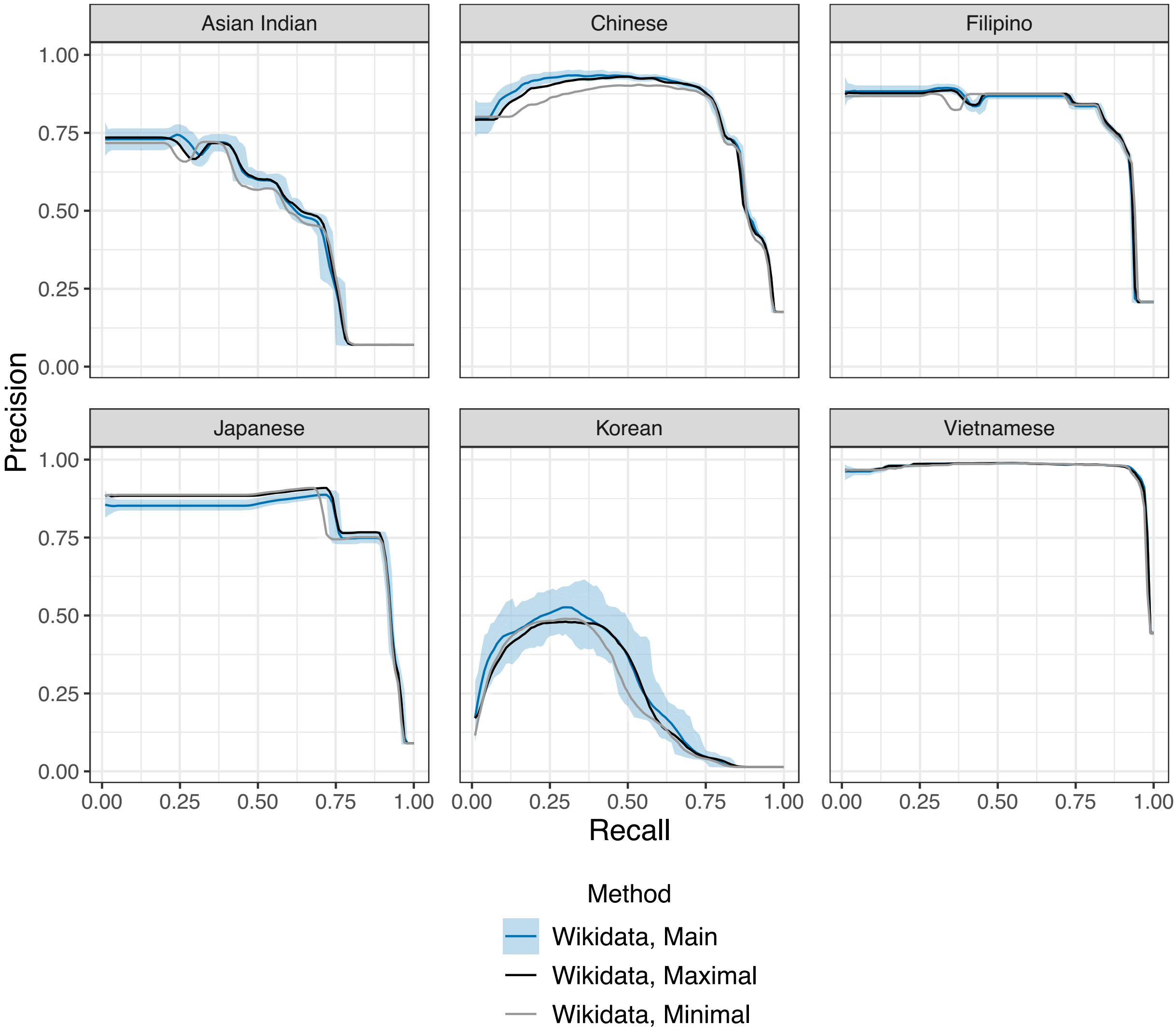

Fig. 6 Precision-Recall Curves, comparing probabilistic name information sourced from Wikidata and processed using the main method (blue), and probabilistic name information sourced from Wikidata and processed using "maximal" (black) and "minimal" (gray) approaches. The shaded region denotes the 95% confidence interval based on 100 bootstrapped PR curves. Intervals for the maximal and minimal methods are omitted for legibility.

cohorts may not be representative of the broader population yet are equally valid candidates for subgroup prediction. That being said, we can also reproduce results from our individual-level and group-level predictions using an alternative validation set. We sample with replacement from the existing validation set to produce a new cohort (including 100 bootstrap iterations for producing confidence intervals) of 20,000 patients that exactly match the overall U.S. proportions of these Asian subgroups (5,490 Asian Indian, 5,619 Chinese, 3,706 Filipino, 958 Japanese, 1,860 Korean, and 2,367 Vietnamese). While the patients used to create this synthetic sample retain the same potential non-representativeness of names and geographies as the original validation set, we expect this revised set to correct for some issues stemming from the imbalanced proportions of subgroups.

As seen in Figs. 4, 5, the revised results differ from the main results in degrees, but generally retain the same takeaways.

Alternative specifications. *Querying.* We explored filtering our query to different ranges of date of birth, but found that a longer time range (in our case, from 1800 to 2020) generally improved performance in our particular validation. We also found that adding descendants to our query contributed about 15% more unique individuals, but did not appreciably improve performance. Lastly, as previously noted, our choices of countries to query as a proxy for each Asian subgroup were intentionally conservative, and could be easily adjusted by users.

Name processing. As shown in Fig. 6, neither the minimal nor maximal approach to name processing achieves better overall performance than our main approach, suggesting that our technique of selecting the best candidate

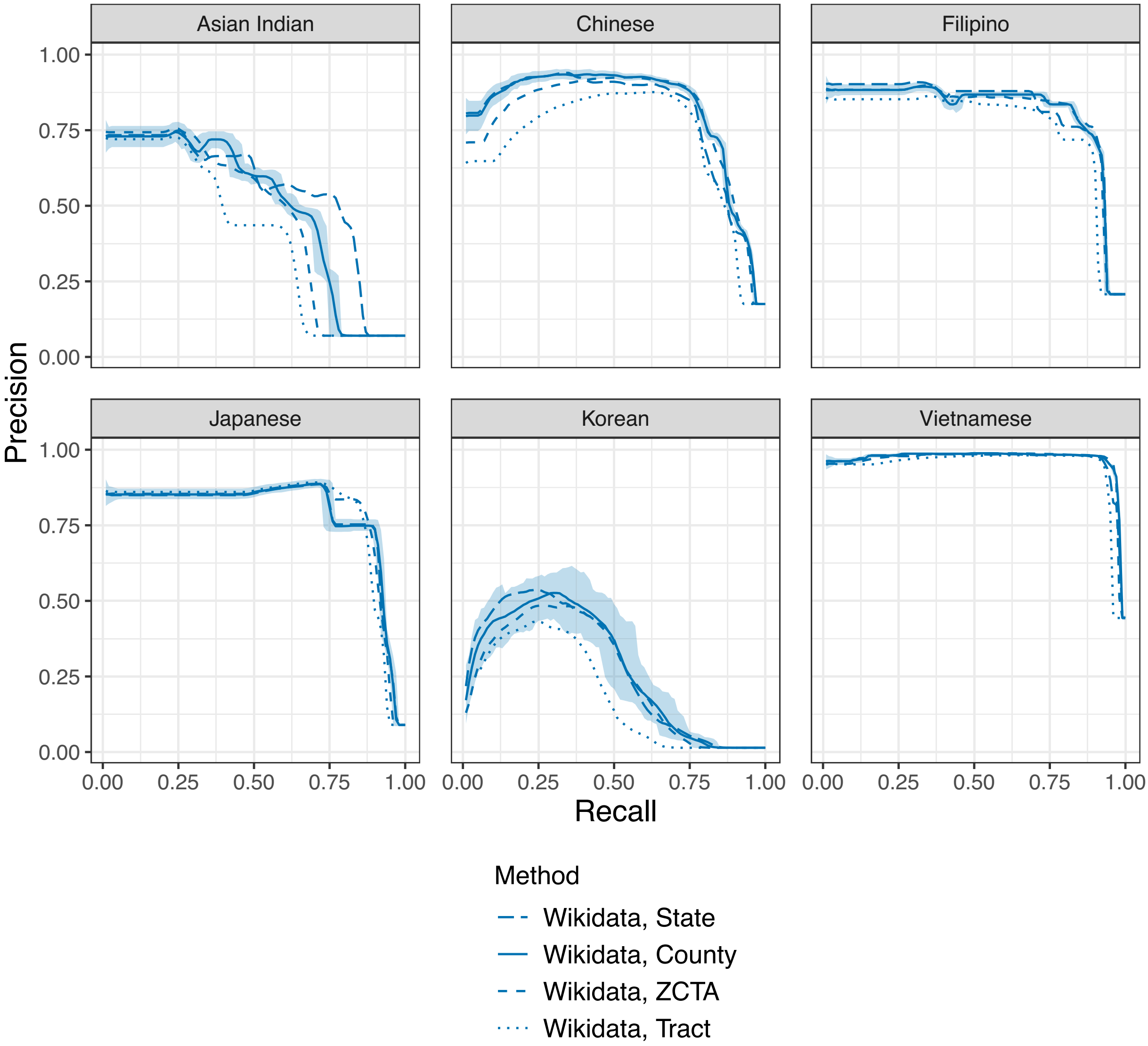

Fig. 7 Precision-Recall Curves, comparing probabilistic name information sourced from Wikidata and implemented using county-level geography information as in the main approach (solid), and alternative implementations using state-level (longdash), ZIP Code tabulation area-level (ZCTA; dashed), and tract-level (dotted) geography information. The shaded region denotes the 95% confidence interval based on 100 bootstrapped PR curves. Intervals for the alternative geographies are omitted for legibility.

names from the unstructured full name field, after having made use of available structured name fields, achieves a useful balance of discovery and curation. However, users may consider the range of approaches described in the Methods section, including alternative means of ranking and selecting one or more plausible candidate names for each individual, for different applications. Our code repository includes the raw outputs of the querying step, so that users can easily explore alternative approaches.

Name algorithm implementation. As shown in Fig. 7, the county-level approach for geography appears most effective overall in our validation, but users are encouraged to test other geography options, which we have provided in our released dataset, for different applications.

Figure 8 shows the PR curves for predictions using both first names and surnames from Wikidata as inputs to BIFSG (blue, dashed), alongside those using just surnames from Wikidata as inputs to BISG (blue, solid) as shown in the manuscript. It also shows the SSA surname approach (orange, solid) shown in the manuscript alongside a variation that incorporates deterministic first name information after the surname matches have already been considered (orange, dashed). The two SSA variations are relatively comparable, while the BIFSG implementation significantly underperforms the BISG implementation for all subgroups. This raises a question about the value of incorporating first names from Wikidata wholesale into imputations of Asian subgroup membership, though researchers may find more specific settings in which they do provide marginal value.

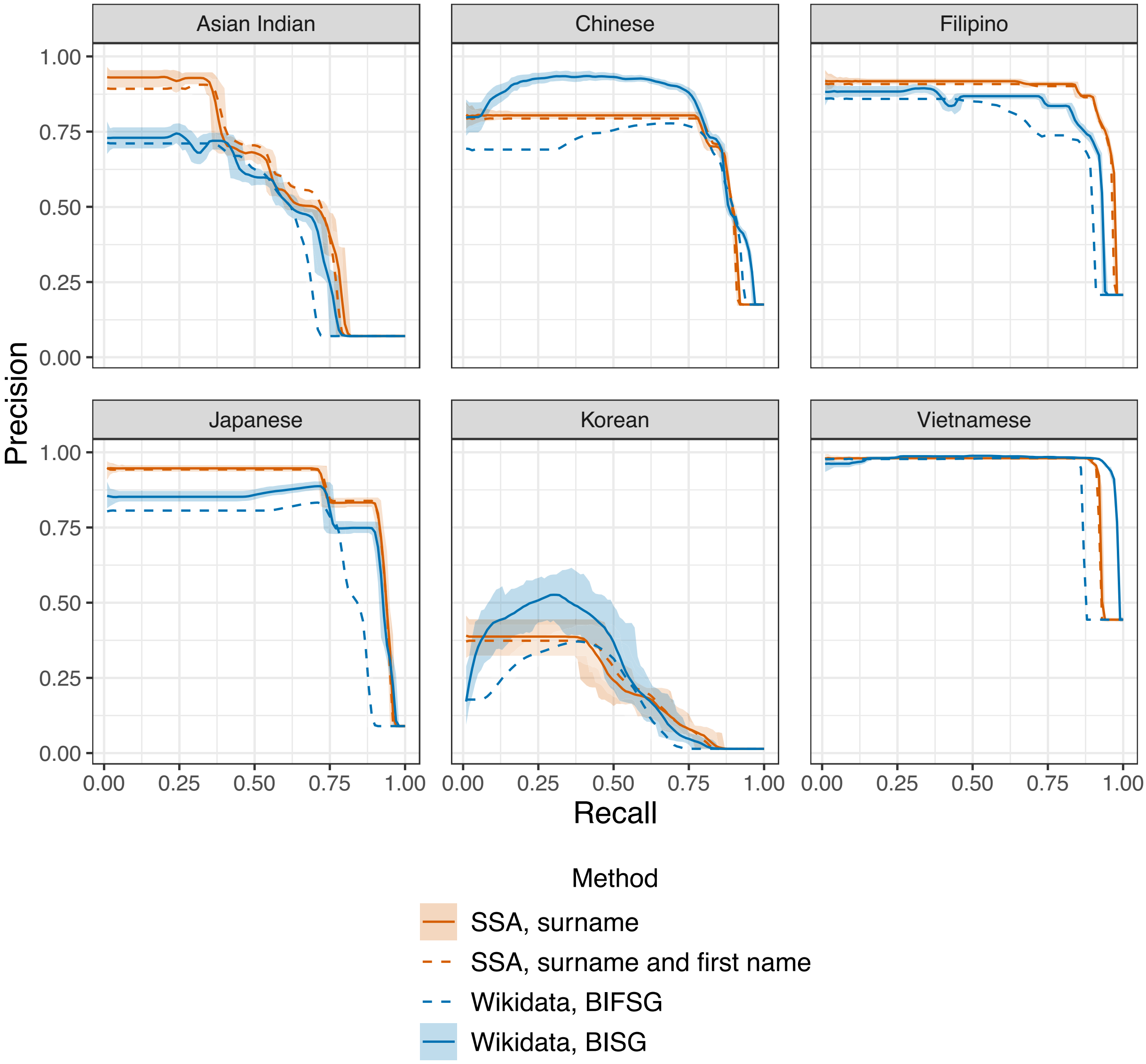

Fig. 8 Precision-Recall Curves, comparing deterministic surname information sourced from SSA (orange, solid), deterministic surname and first name information sourced from SSA (orange, dashed), probabilistic surname information sourced from Wikidata and used in BISG (blue, solid), and probabilistic surname and first name information sourced from Wikidata and used in BIFSG (blue, dashed). The shaded region denotes the 95% confidence interval based on 100 bootstrapped PR curves. Intervals for the first name methods are omitted for legibility.

Subgroup-level health disparity prediction. We experimented with various alternative approaches for estimating subgroup-level prevalences and a variety of parameter settings, but ultimately could not find any which consistently performed better than Eq. 2. One key assumption underlying the techniques that use only a subset of individuals is that, conditional on predicted race, there is no residual correlation between the outcome and actual race; this is the same assumption we underscored in the main approach. As we found in our evaluation, this may be far from true, in which case restricting to a subset of individuals from one end of the range of posterior probabilities may further exacerbate associated biases. That being said, users may find that this assumption is less violated in different settings, which may lead to some of these alternative approaches performing better for subgroup-level estimation.

Historical census names. Fig. 9 shows the PR curves for predictions using names from historical censuses, alongside those using names from Wikidata. Both approaches make use of the same geography-based priors, as appropriate.

While predictions using historical census names do achieve a higher precision within a certain range of recall for Filipino, Japanese, and Korean predictions, in virtually all other cases, they substantially and significantly underperform both the Wikidata approach. One key reason for this is that the historical censuses we have access

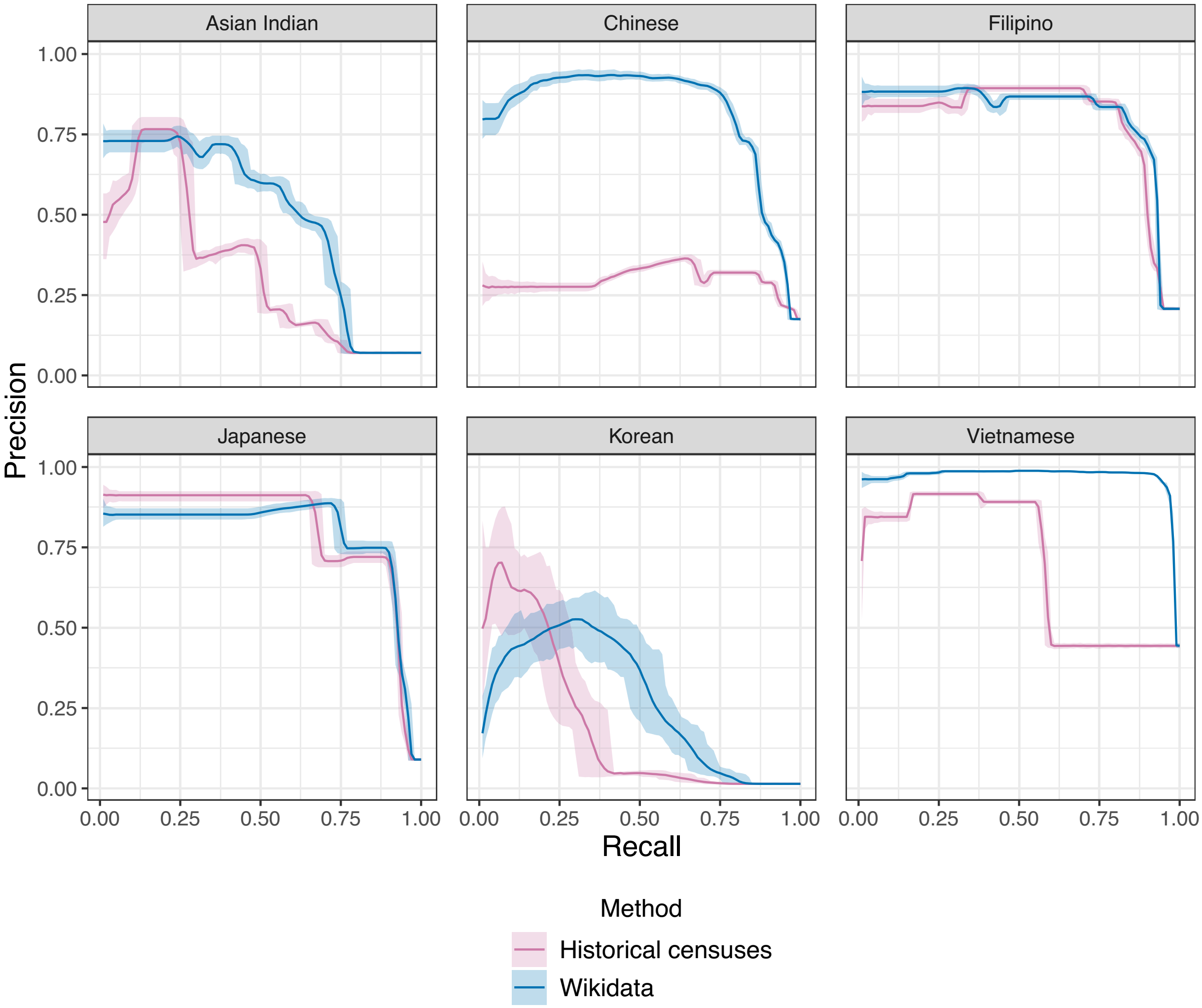

Fig. 9 Precision-Recall Curves, comparing probabilistic name information sourced from Wikidata (blue), and probabilistic name information sourced from historical censuses (pink). The shaded region denotes the 95% confidence interval based on 100 bootstrapped PR curves.

to predate the primary waves of Vietnamese immigration beginning in the 1970s, and thus did not include any individuals identified specifically as Vietnamese[88], significantly impacting performance for the Vietnamese subgroup (which can only use geography-based predictions) and other subgroups (since actual Vietnamese individuals are more likely to be erroneously classified as members of other subgroups).

Usage Notes

Users are encouraged to review our published codebase for further documentation and guidance. In particular, users should perform similar steps of string cleaning on their names (using the functions provided in our codebase) as we perform on Wikidata names, and, if working with raw address information, must perform their own geocoding to map individuals to census geographies. Given clean inputs of name and census geography for individuals, users can then directly use their records as inputs, along with our published name-race and geography-race distribution tables, to BISG and BIFSG name algorithms, for which we provide a convenient function in our codebase. Users are also encouraged to review the Methods and Technical Validation sections in this manuscript for guidance on how to utilize a subset of individuals with reported subgroup information, if available, to calibrate predictions, given the variety of alternative specifications which may exhibit better performance in particular settings. To utilize the hybrid approach, combining names from Wikidata with names from SSA, users should submit an email request to the authors of the complete SSA name lists[13], and then may use the SSA names as inputs in our code. This technique may also be used to test hybrid approaches combining any two or more sources of name information.

Limitations. Our dataset consists of name-race distributions for Asian subgroups derived from Wikidata, which are designed to be used as inputs to BISG or BIFSG in settings where individuals are already known to be Asian, and the objective is to predict their membership in six major Asian subgroups. However, these six major subgroups only comprise 80% of the total Asian population in the U.S., per ACS 2022 estimates[76]. Individuals who identify with another Asian subgroup (*e.g*., Thai, Hmong, etc.) would receive an incorrect prediction. In the Technical Validation section, we focus on evaluating performance on a restricted cohort who only self-report as one of the six major Asian subgroups. Note that these six subgroups are also the subgroups covered by the most commonly used existing name lists[13] and required as disaggregated options in the new 2024 standards for federal data collection.

Our EHR validation dataset, while broadly representative of patients across the U.S., is not without its limitations. First, race may be recorded erroneously for multiple reasons as documented in prior studies[89], affecting the EHR data's reliability for validation. 81% of patients in our validation set, who are recorded as one of the six subgroups, also would be predicted as being API using conventional BISG. However, this alignment rate varies substantially by subgroup, from a high of 97% for Vietnamese to a low of 41% for Asian Indian. In the latter group, the posterior probability from conventional BISG of being AIAN is notably high, suggesting word ambiguity (*e.g*., "Indian") in the race recording process as a source of discordance that may distort some of our validation results. Second, as previously noted, fewer than 15% of patients who self-identify as Asian also provide a detailed subgroup membership, and such patients may differ from the broader population of Asian Indian, Chinese, Filipino, Japanese, Korean, and Vietnamese individuals in terms of names, locations, and health outcomes. For instance, the median age of patients in our validation set is 58, compared to 38 for Asian Americans nationwide. That being said, our methodology is broadly applicable in any setting in which users are interested in assessing group-level disparities across all available individuals, but only have recorded subgroup information for a small subset (or none at all, in which case the option may exist to expend resources to obtain race information for a small subset). We provide a reproducible and flexible framework for evaluating accuracy on individual-level and group-level prediction tasks across a wide array of data collection, data processing, and model specification choices, as detailed in the Methods section. Given whatever variation is observed in performance on the test subset, users can then select the version of our method that appears most calibrated to their particular setting and, with confidence appropriately bounded by our uncertainty quantification technique, proceed to perform race imputation more widely.

Despite the promising results shown in this work, Wikidata is, of course, an imperfect data source given its biases, errors, and incompleteness[90,91]. Recent work suggests that it is possible to reduce errors by cross-checking different language editions of Wikipedia[92]. One of the practical benefits of utilizing an open, crowdsourced knowledge base like Wikidata is that it can be expected to grow in coverage and improve in detail over time. As Wikidata grows, more names can be incorporated into our name-race distribution tables, likely increasing the percentage of names that are observed across more than one subgroup and that receive the benefits of probabilistic information demonstrated in this work.

Code availability

The online repository for replication code and intermediate results is on GitHub (https://github.com/reglab/disaggregation).

Received: 11 November 2024; Accepted: 4 March 2025;
Published: 5 April 2025

References

1. The Release of the Equitable Data Working Group Report | OSTP [Internet]. The White House. [cited 2024 May 30]. Available from: https://www.whitehouse.gov/ostp/news-updates/2022/04/22/the-release-of-the-equitable-data-working-group-report/ (2022).
2. Chin, M. K. *et al*. Methods for Retrospectively Improving Race/Ethnicity Data Quality: A Scoping Review. *Epidemiol Rev [Internet]*. 2023 Apr 12 [cited 2023 Nov 20];mxad002. Available from: https://doi.org/10.1093/epirev/mxad002.
3. Elliott, M. N. *et al*. Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Serv Outcomes Res Methodol [Internet]*. [cited 2023 Nov 20]; **9**(2), 69–83. Available from: https://doi.org/10.1007/s10742-009-0047-1 (2009).
4. Imai, K. & Khanna, K. Improving ecological inference by predicting individual ethnicity from voter registration records. *Polit Anal*. **24**(2), 263–72 (2016).
5. Zhang, Y. Assessing Fair Lending Risks Using Race/Ethnicity Proxies. *Manag Sci [Internet]*. [cited 2023 Nov 22]; **64**(1), 178–97. Available from: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.2016.2579 (2018).
6. Greenwald, D., Howell, S. T., Li, C. & Yimfor, E. Regulatory Arbitrage or Random Errors? Implications of Race Prediction Algorithms in Fair Lending Analysis [Internet]. *National Bureau of Economic Research*; [cited 2023 Nov 22]. (Working Paper Series). Available from: https://www.nber.org/papers/w31646 (2023).
7. The Consumer Financial Protection Bureau. Using Publicly Available Information to Proxy for Unidentified Race and Ethnicity: A Methodology and Assessment [Internet]. *The Consumer Financial Protection Bureau*; [cited 2024 Jun 12]. Available from: https://files.consumerfinance.gov/f/201409_cfpb_report_proxy-methodology.pdf (2014).
8. Fraga, B. L. *The Turnout Gap: Race, Ethnicity, and Political Inequality in a Diversifying America [Internet]*. Cambridge: Cambridge University Press; [cited 2023 Nov 20]. Available from: https://www.cambridge.org/core/books/turnout-gap/1B79B19C880A93C462FD1DF22F65DD15 (2018).
9. Decter-Frain, A. *et al*. Comparing Methods for Estimating Demographics in Racially Polarized Voting Analyses. *Sociol Methods Res [Internet]*. [cited 2023 Nov 20]; 00491241231192383. Available from: https://doi.org/10.1177/00491241231192383 (2023).
10. Hepburn, P., Louis, R. & Desmond, M. Racial and Gender Disparities among Evicted Americans. *Sociol Sci [Internet]*. [cited 2023 Nov 20]; **7**, 649–62. Available from: https://sociologicalscience.com/articles-v7-27-649/ (2020).
11. Colorado Division of Insurance. DRAFT PROPOSED Algorithm and Predictive Model Quantitative Testing Regulation [Internet]. Colorado Division of Insurance; [cited 2024 Jun 12]. Available from: https://drive.google.com/file/d/1BMFuRKbh39Q7YckPqrhrCRuWp29vJ44O/view (2023).
12. Elzayn, H. *et al*. Measuring and Mitigating Racial Disparities in Tax Audits. *Q J Econ*. 2024 forthcoming;
13. Lauderdale, D. S. & Kestenbaum, B. Asian American ethnic identification by surname. *Popul Res Policy Rev [Internet]*. [cited 2023 Nov 14]; **19**(3), 283–300. Available from: https://doi.org/10.1023/A:1026582308352 (2000).

14. The White House. *Executive Order On Advancing Racial Equity and Support for Underserved Communities Through the Federal Government [Internet]*. Available from: https://www.whitehouse.gov/briefing-room/presidential-actions/2021/01/20/executive-order-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/ (2021).
15. The White House. *Executive Order on Further Advancing Racial Equity and Support for Underserved Communities Through The Federal Government [Internet]*. Available from: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/ (2023).
16. Office of Management and Budget. Revisions to OMB's Statistical Policy Directive No. 15: Standards for Maintaining, Collecting, and Presenting Federal Data on Race and Ethnicity [Internet]. [cited 2024 May 29]. Available from: https://www.federalregister.gov/documents/2024/03/29/2024-06469/revisions-to-ombs-statistical-policy-directive-no-15-standards-for-maintaining-collecting-and (2024).
17. The Privacy Act of 1974 [Internet]. 5 U.S.C. § 552a 1974. Available from: https://epic.org/the-privacy-act-of-1974/.
18. Andrus, M., Spitzer, E., Brown, J. & Xiang, A. What We Can't Measure, We Can't Understand: Challenges to Demographic Data Procurement in the Pursuit of Fairness. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. **2021**. p. 249–60.
19. King, J., Ho, D., Gupta, A., Wu, V. & Webley-Brown, H. The Privacy-Bias Tradeoff: Data Minimization and Racial Disparity Assessments in U.S. Government. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery; 2023 [cited 2023 Aug 15]*. p. 492–505. (FAccT '23). Available from: https://dl.acm.org/doi/10.1145/3593013.3594015.
20. Office of Management and Budget. *Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity [Internet]*. [cited 2024 Jan 30]. Available from: https://obamawhitehouse.archives.gov/omb/fedreg_1997standards (1997)
21. Comenetz, J. *Frequently Occuring Surnames in the 2010 Census [Internet]. the Census Bureau; 2016 [cited 2023 Nov 10]*. https://www2.census.gov/topics/genealogy/2010surnames/surnames.pdf.
22. Kauh, T. J., Read, J. G. & Scheitler, A. J. The Critical Role of Racial/Ethnic Data Disaggregation for Health Equity. *Popul Res Policy Rev [Internet]*. [cited 2022 Oct 20]; **40**(1), 1–7. Available from: http://link.springer.com/10.1007/s11113-020-09631-6 (2021).
23. Shimkhada, R., Scheitler, A. J. & Ponce, N. A. Capturing Racial/Ethnic Diversity in Population-Based Surveys: Data Disaggregation of Health Data for Asian American, Native Hawaiian, and Pacific Islanders (AANHPIs). *Popul Res Policy Rev [Internet]*. [cited 2024 Jun 3]; **40**(1), 81–102. Available from: https://doi.org/10.1007/s11113-020-09634-3 (2021).
24. Panapasa, S. V., Crabbe, K. M. & Kaholokula, J. K. Efficacy of Federal Data: Revised Office of Management and Budget Standard for Native Hawaiian and Other Pacific Islanders Examined. *AAPI Nexus Asian Am Pac Isl Policy Pract Community [Internet]*. [cited 2024 Jun 14]; **9**(1–2), 212–20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4211287/ (2011).
25. Llanos, K. & Palmer, L. *Using Data on Race and Ethnicity to Improve Health Care Quality for Medicaid Beneficiaries [Internet]. Center for Health Care Strategies, Incorporated;* [cited 2024 Jun 10]. Available from: https://www.chcs.org/media/Using_Date_to_Reduce_Health_Disparities.pdf (2006).
26. Chang, R. C., Penaia, C. & Thomas, K. Count Native Hawaiian And Pacific Islanders In COVID-19 Data—It's An OMB Mandate. *Health Aff Forefr [Internet]*. [cited 2023 Nov 20]; Available from: https://www.healthaffairs.org/do/10.1377/forefront.20200825.671245/full/ (2020).
27. *Equal Credit Opportunity Act [Internet]. 15 U.S.C. §§ 1691-1691f 1974*. Available from: https://www.ecfr.gov/current/title-12/chapter-X/part-1002/subpart-A/section-1002.5.
28. *Protecting Consumers from Unfair Discrimination in Insurance Practices [Internet]*. 10-3-1104.9 2021. Available from: https://leg.colorado.gov/bills/sb21-169.
29. Klein, D. J. *et al.* Understanding Nonresponse to the 2007 Medicare CAHPS Survey. *The Gerontologist [Internet]*. [cited 2024 Jun 3]; **51**(6), 843–55. Available from: https://doi.org/10.1093/geront/gnr046 (2011).
30. Islam, N. S. *et al.* Methodological Issues in the Collection, Analysis, and Reporting of Granular Data in Asian American Populations: Historical Challenges and Potential Solutions. *J Health Care Poor Underserved [Internet]*. [cited 2024 Jun 3]; **21**(4), 1354–81. Available from: https://muse.jhu.edu/pub/1/article/400774 (2010).
31. Wang, H. L. *"Racist Bill"? Chinese Immigrants Protest Effort To Collect More Asian-American Data*. [cited 2024 Jun 1]; Available from: https://www.npr.org/2017/08/05/541844705/protests-against-the-push-to-disaggragate-asian-american-data (2017).
32. Darity, W. A. & Lefebvre, S. Data Collection without Definitions: Problems with OMB Directive 15 and a Proposal. In: *Race, Ethnicity, and Economic Statistics for the 21st Century [Internet]. University of Chicago Press*; [cited 2024 Jun 17]. Available from: https://www.nber.org/books-and-chapters/race-ethnicity-and-economic-statistics-21st-century/data-collection-without-definitions-problems-omb-directive-15-and-proposal (2024).
33. Srinivasan, S. & Guillermo, T. Toward improved health: disaggregating Asian American and Native Hawaiian/Pacific Islander data. *Am J Public Health*. **90**(11), 1731 (2000).
34. Holland, A. T. & Palaniappan, L. P. Problems With the Collection and Interpretation of Asian-American Health Data: Omission, Aggregation, and Extrapolation. *Ann Epidemiol [Internet]*. [cited 2023 Aug 28]; **22**(6), 397–405. Available from: https://www.sciencedirect.com/science/article/pii/S1047279712000956 (2012).
35. Yom, S. & Lor, M. Advancing Health Disparities Research: The Need to Include Asian American Subgroup Populations. *J Racial Ethn Health Disparities [Internet]*. [cited 2023 Nov 26]; **9**(6), 2248–82. Available from: https://doi.org/10.1007/s40615-021-01164-8 (2022).
36. Kibria, N. The contested meanings of "Asian American": racial dilemmas in the contemporary US. *Ethn Racial Stud [Internet]*. [cited 2024 Jun 11]; **21**(5), 939–58. Available from: https://doi.org/10.1080/014198798329739 (1998).
37. Kim, J. H. J., Lu, Q. & Stanton, A. L. Overcoming constraints of the model minority stereotype to advance Asian American health. *Am Psychol*. **76**(4), 611–26 (2021).
38. Kuo, E. E., Kraus, M. W. & Richeson, J. A. High-Status Exemplars and the Misperception of the Asian-White Wealth Gap. *Soc Psychol Personal Sci [Internet]*. [cited 2024 Jun 5]; **11**(3), 397–405. Available from: https://doi.org/10.1177/1948550619867940 (2020).
39. Akee, R., Jones, M. R. & Porter, S. R. Race Matters: Income Shares, Income Inequality, and Income Mobility for All U.S. Races. *Demography [Internet]*. [cited 2023 Nov 20]; **56**(3), 999–1021. Available from: https://doi.org/10.1007/s13524-019-00773-7 (2019).
40. Vu, M. *et al.* Low-Income Asian Americans: High Levels Of Food Insecurity And Low Participation In The CalFresh Nutrition Program. *Health Aff (Millwood) [Internet]*. [cited 2023 Nov 20]; **42**(10), 1420–30. Available from: https://www.healthaffairs.org/doi/10.1377/hlthaff.2023.00116 (2023).
41. Budiman, A. & Ruiz, N. G. *Key facts about Asian origin groups in the U.S. [Internet]*. Pew Research Center; [cited 2023 Nov 27]. Available from: https://www.pewresearch.org/short-reads/2021/04/29/key-facts-about-asian-origin-groups-in-the-u-s/ (2021).
42. Acciai, F., Noah, A. J. & Firebaugh, G. Pinpointing the sources of the Asian mortality advantage in the USA. *J Epidemiol Community Health [Internet]*. [cited 2024 May 31]; **69**(10), 1006–11. Available from: https://jech.bmj.com/lookup/doi/10.1136/jech-2015-205623 (2015).
43. Baluran, D. A. & Patterson, E. J. Examining Ethnic Variation in Life Expectancy Among Asians in the United States, 2012–2016. *Demography [Internet]*. [cited 2023 Aug 30]; **58**(5), 1631–54. Available from: https://doi.org/10.1215/00703370-9429449 (2021).
44. New UCSF Study to Find out What Drives Cancer in Asian Americans | UC San Francisco [Internet]. [cited 2024 Jun 11]. Available from: https://www.ucsf.edu/news/2024/05/427586/new-ucsf-study-find-out-what-drives-cancer-asian-americans (2024).
45. Shah, N. S. *et al.* Heterogeneity in Obesity Prevalence Among Asian American Adults. *Ann Intern Med [Internet]*. [cited 2023 Nov 21]; **175**(11), 1493–500. Available from: https://www.acpjournals.org/doi/10.7326/M22-0609 (2022).

46. Shah, N. S. *et al*. Self-Reported Diabetes Prevalence in Asian American Subgroups: Behavioral Risk Factor Surveillance System, 2013–2019. *J Gen Intern Med [Internet]*. [cited 2023 Nov 28]; **37**(8), 1902–9. Available from: https://doi.org/10.1007/s11606-021-06909-z (2022).
47. Torre, L. A. *et al*. Cancer statistics for Asian Americans, Native Hawaiians, and Pacific Islanders, 2016: Converging incidence in males and females. *CA Cancer J Clin [Internet]*. [cited 2023 Nov 22]; **66**(3), 182–202. Available from: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21335 (2016).
48. Kanaya, A. M. *et al*. Knowledge Gaps, Challenges, and Opportunities in Health and Prevention Research for Asian Americans, Native Hawaiians, and Pacific Islanders: A Report From the 2021 National Institutes of Health Workshop. *Ann Intern Med [Internet]*. [cited 2023 Sep 11]; **175**(4), 574–89. Available from: https://www.acpjournals.org/doi/10.7326/M21-3729 (2022).
49. Kalyanaraman Marcello, R. *et al*. Disaggregating Asian Race Reveals COVID-19 Disparities Among Asian American Patients at New York City's Public Hospital System. *Public Health Rep [Internet]*. Mar [cited 2022 Oct 20]; **137**(2), 317–25. Available from: http://journals.sagepub.com/doi/10.1177/00333549211061313 (2022).
50. Schwartz, G. L. & Jahn, J. L. Disaggregating Asian American and Pacific Islander Risk of Fatal Police Violence. *PLOS ONE [Internet]*. [cited 2023 Aug 28]; **17**(10), e0274745. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0274745 (2022).
51. Wong, E. C., Palaniappan, L. P. & Lauderdale, D. S. Using Name Lists to Infer Asian Racial/Ethnic Subgroups in the Healthcare Setting. *Med Care [Internet]*. [cited 2024 May 28]; **48**(6), 540. Available from: https://journals.lww.com/lww-medicalcare/fulltext/2010/06000/Using_Name_Lists_to_Infer_Asian_Racial_Ethnic.00008.aspx (2010).
52. Kozlowski, D. *et al*. Avoiding bias when inferring race using name-based approaches. *PLOS ONE [Internet]*. [cited 2023 Nov 22]; **17**(3), e0264270. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264270 (2022).
53. Dong, E., Schein, A., Wang, Y. & Garg, N. *Addressing Discretization-Induced Bias in Demographic Prediction [Internet]*. arXiv; [cited 2024 May 30]. Available from: http://arxiv.org/abs/2405.16762 (2024).
54. Chen, J., Kallus, N., Mao, X., Svacha, G. & Udell, M. Fairness under unawareness: Assessing disparity when protected class is unobserved. In: *Proceedings of the conference on fairness, accountability, and transparency*. p. 339–48 (2019).
55. Elzayn, H. *et al*. Estimating and Implementing Conventional Fairness Metrics With Probabilistic Protected Features. In: *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML) [Internet]*. Toronto, ON, Canada: IEEE; [cited 2024 Sep 27]. p. 161–93. Available from: https://ieeexplore.ieee.org/document/10516629/ (2024).
56. Adjaye-Gbewonyo, D., Bednarczyk, R. A., Davis, R. L. & Omer, S. B. 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 Serv Res*. **49**(1), 268–83 (2014).
57. Louie, E. W. *Chinese American Names: Tradition and Transition. McFarland*; 239 p (2008).
58. Fryer, R. Guess Who's Been Coming to Dinner? Trends in Interracial Marriage over the 20th Century. *J Econ Perspect [Internet]*. [cited 2024 Jun 3]; **21**(2), 71–90. Available from: https://www.aeaweb.org/articles?id=10.1257/jep.21.2.71 (2007).
59. Gullickson, A. Patterns of Panethnic Intermarriage in the United States, 1980–2018. *Demography [Internet]*. [cited 2024 May 29]; **59**(5), 1929–51. Available from: https://doi.org/10.1215/00703370-10218826 (2022).
60. Tzioumis, K. Demographic aspects of first names. *Sci Data*. **5**(1), 1–9 (2018).
61. Rosenman, E. T. R., Olivella, S. & Imai, K. Race and ethnicity data for first, middle, and surnames. *Sci Data [Internet]*. [cited 2023 Aug 28]; **10**(1), 299. Available from: https://www.nature.com/articles/s41597-023-02202-2 (2023).
62. Vrandečić, D. & Krötzsch, M. Wikidata: a free collaborative knowledgebase. *Commun ACM [Internet]*. [cited 2023 Oct 4]; **57**(10), 78–85. Available from: https://dl.acm.org/doi/10.1145/2629489 (2014).
63. Poston, D. L. & Wong, J. H. The Chinese diaspora: The current distribution of the overseas Chinese population. *Chin J Sociol [Internet]*. [cited 2023 Nov 21]; **2**(3), 348–73. Available from: https://doi.org/10.1177/2057150X16655077 (2016).
64. Cheng, L., Gallegos, I. O., Ouyang, D., Goldin, J. & Ho, D. How Redundant are Redundant Encodings? Blindness in the Wild and Racial Disparity when Race is Unobserved. In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency [Internet]. New York, NY, USA: Association for Computing Machinery*; [cited 2023 Nov 28]. p. 667–86. (FAccT '23). Available from: https://dl.acm.org/doi/10.1145/3593013.3594034 (2023).
65. Kulke, H. & Rothermund, D. *A History of India [Internet]. 0 ed. Routledge*; [cited 2025 Feb 10]. Available from: https://www.taylorfrancis.com/books/9781317242130 (2016).
66. Grasso, J., Corrin, J. P. & Kort, M. *Modernization and Revolution in China [Internet]*. 5th ed. Routledge; [cited 2025 Feb 10]. Available from: https://www.taylorfrancis.com/books/9781317236641 (2017).
67. Nadeau, K. *The history of the Philippines*. Bloomsbury Publishing USA; (2020).
68. Hane, M. & Perez, L. G. *Modern Japan: A Historical Survey [Internet]. 5th ed*. Routledge; [cited 2025 Feb 10]. Available from: https://www.taylorfrancis.com/books/9780429963520 (2018).
69. Buzo, A. *The Making of Modern Korea [Internet]. 4th ed*. London: Routledge; [cited 2025 Feb 10]. Available from: https://www.taylorfrancis.com/books/9781003241706 (2022).
70. Karnow, S. *Vietnam: A history*. **Vol. 122**. Random House; (1994).
71. Palumbo-Liu, D. *Asian/American: Historical crossings of a racial frontier*. Stanford University Press; (1999).
72. Clark, J. T., Curiel, J. A. & Steelman, T. S. Minmaxing of Bayesian Improved Surname Geocoding and Geography Level Ups in Predicting Race. *Polit Anal [Internet]*. [cited 2023 Nov 24]; **30**(3), 456–62. Available from: https://www.cambridge.org/core/journals/political-analysis/article/minmaxing-of-bayesian-improved-surname-geocoding-and-geography-level-ups-in-predicting-race/2B259C0A8B66EFB00C4AD05B19CCFF4A (2022).
73. Imai, K., Olivella, S. & Rosenman, E. T. R. Addressing census data problems in race imputation via fully Bayesian Improved Surname Geocoding and name supplements. *Sci Adv [Internet]*.; **8**(49). Available from: https://pubmed.ncbi.nlm.nih.gov/36490334/ (2022).
74. Lu, B., Wan, J., Ouyang, D., Goldin, J. & Ho, D. E. *Quantifying the Uncertainty of Imputed Demographic Disparity Estimates: The Dual-Bootstrap [Internet]*. National Bureau of Economic Research; [cited 2024 Apr 14]. (Working Paper Series). Available from: https://www.nber.org/papers/w32312 (2024).
75. Voicu, I. Using First Name Information to Improve Race and Ethnicity Classification. *Stat Public Policy [Internet]*. [cited 2022 Oct 18]; **5**(1), 1–13. Available from: https://doi.org/10.1080/2330443X.2018.1427012 (2018).
76. Ruggles, S. *et al*. *IPUMS USA: Version 15.0 [Internet]*. Minneapolis, MN: IPUMS; [cited 2024 Aug 16]. Available from: https://usa.ipums.org (2024).
77. Lin, Q. *et al*. Enabling Disaggregation of Asian American Subgroups: A Dataset of Wikidata Names for Disparity Estimation [Internet]. *Harvard Dataverse*; [cited 2024 Oct 31]. Available from: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/LEOECM (2024).
78. Vala, A., Hao, S., Chu, I., Phillips, R. L. & Rehkopf, D. *The American Family Cohort (v12. 2). Redivis [Internet]*. Stanford, CA.; Available from: https://doi.org/10.57761/jn2e-7r28 (2023).
79. Lu, L. *et al*. A language-matching model to improve equity and efficiency of COVID-19 contact tracing. *Proc Natl Acad Sci*. **118**(43) (2021).
80. Lu, L., D'Agostino, A., Rudman, S. L., Ouyang, D. & Ho, D. E. Designing Accountable Healthcare Algorithms: A Case Study from COVID-19 Contact Tracing. *N Engl J Med Catal*. (2022).

81. Chugg, B. *et al*. Evaluation of allocation schemes of COVID-19 testing resources in a community-based door-to-door testing program. *In: JAMA Health Forum*. p. e212260–e212260 American Medical Association (2021).
82. Office of Minority Health. *Asian American Health [Internet]*. [cited 2023 Nov 20]. Available from: https://minorityhealth.hhs.gov/asian-american-health (2023).
83. Reich, C. *et al*. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *J Am Med Inform Assoc [Internet]*. [cited 2024 Aug 16]; **31**(3), 583–90. Available from: https://academic.oup.com/jamia/article/31/3/583/7510741 (2024).
84. Ford, C. N. *et al*. Racial differences in performance of HbA$_{1c}$ for the classification of diabetes and prediabetes among US adults of non-Hispanic black and white race. *Diabet Med [Internet]*. [cited 2024 Oct 28]; **36**(10), 1234–42. Available from: https://onlinelibrary.wiley.com/doi/10.1111/dme.13979 (2019).
85. Hahm, H. C., Cook, B. L., Ault-Brutus, A. & Alegría, M. Intersection of Race-Ethnicity and Gender in Depression Care: Screening, Access, and Minimally Adequate Treatment. *Psychiatr Serv [Internet]*. [cited 2024 Oct 28]; **66**(3), 258–64. Available from: https://psychiatryonline.org/doi/10.1176/appi.ps.201400116 (2015).
86. Wang, A., Ramaswamy, V. V. & Russakovsky, O. Towards Intersectionality in Machine Learning: Including More Identities, Handling Underrepresentation, and Performing Evaluation. In: *2022 ACM Conference on Fairness, Accountability, and Transparency [Internet]*. Seoul Republic of Korea: ACM; [cited 2024 Sep 24]. p. 336–49. Available from: https://dl.acm.org/doi/10.1145/3531146.3533101 (2022).
87. Kobayashi, K., Chan, K. T. K., Roy, A., Khan, M. M. & Fuller-Thomson, E. Diabetes and Diabetes Care among Nonobese Japanese-Americans: Findings from a Population-Based Study. *Adv Prev Med [Internet]*. [cited 2024 Oct 28]; **2019**, 1–8. Available from: https://www.hindawi.com/journals/apm/2019/3650649/ (2019).
88. Humes, K. & Hogan, H. Measurement of Race and Ethnicity in a Changing, Multicultural America. *Race Soc Probl [Internet]*. [cited 2024 Aug 20]; **1**(3), 111–31. Available from: http://link.springer.com/10.1007/s12552-009-9011-5 (2009).
89. Johnson, J. A., Moore, B., Hwang, E. K., Hickner, A. & Yeo, H. The accuracy of race & ethnicity data in US based healthcare databases: A systematic review. *Am J Surg [Internet]*. [cited 2024 Mar 5]; **226**(4), 463–70. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0002961023001976 (2023).
90. Brown, A. R. Wikipedia as a Data Source for Political Scientists: Accuracy and Completeness of Coverage. *PS Polit Sci Polit [Internet]*. [cited 2023 Nov 20]; **44**(2), 339–43. Available from: https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/wikipedia-as-a-data-source-for-political-scientists-accuracy-and-completeness-of-coverage/DAC48E1EB5C400B92487DADDA63D2216 (2011).
91. Shaik, Z., Ilievski, F. & Morstatter, F. *Analyzing Race and Country of Citizenship Bias in Wikidata [Internet]*. arXiv; 2021 [cited 2023 Sep 7]. arXiv https://arxiv.org/abs/2108.05412 (2021).
92. Laouenan, M. *et al*. A cross-verified database of notable people, 3500BC-2018AD. *Sci Data [Internet]*. [cited 2023 Aug 28]; **9**(1), 290. Available from: https://www.nature.com/articles/s41597-022-01369-4 (2022).

Acknowledgements

We thank Diane Lauderdale for generously providing us with the first and last name lists sourced from the historical Social Security Administration files; the Stanford Center for Population Health Sciences and the American Board of Family Medicine for their expertise on primary care clinical practice; Kit Rodolfa, Jennah Gosciak, Aparna Balagopalan, Jennifer Wang, Benji Lu, Angelina Wang, and Sarfaraz Alam for helpful comments and conversations; and the Stanford Institute for Human-Centered Artificial Intelligence for supporting this work. Isabel O. Gallegos is supported by the Fannie & John Hertz Foundation and Stanford Knight-Hennessy Scholars graduate fellowship.

Author contributions

D.O. and D.E.H. conceived the study; Q.L., D.O., C.G., and I.O.G. developed the methodology; Q.L. and D.O. produced and analyzed the results. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.E.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.