Imputation of race and ethnicity categories using genetic ancestry from real-world genomic testing data

Brooke Rhead[*], Paige E. Haffener[*], Yannick Pouliot, and Francisco M. De La Vega[†]

Tempus Labs, Inc.
Chicago, IL, 60654, USA

The incompleteness of race and ethnicity information in real-world data (RWD) hampers its utility in promoting healthcare equity. This study introduces two methods—one heuristic and the other machine learning-based—to impute race and ethnicity from genetic ancestry using tumor profiling data. Analyzing de-identified data from over 100,000 cancer patients sequenced with the Tempus xT panel, we demonstrate that both methods outperform existing geolocation and surname-based methods, with the machine learning approach achieving high recall (range: 0.859-0.993) and precision (range: 0.932-0.981) across four mutually exclusive race and ethnicity categories. This work presents a novel pathway to enhance RWD utility in studying racial disparities in healthcare.

Keywords: Race; ethnicity; ancestry; imputation, disparities, equity, real-world data.

1. Introduction

Real-world data (RWD) offers insights into disease etiology, therapy outcomes, and racial disparities in healthcare.[1,2] However, its utility in improving healthcare equity is limited by the significant sparsity of race and ethnicity data. This gap, attributable to factors such as lack of capture, data loss during transfer and de-identification,[3,4] and shortcomings in electronic health record integrations,[5] leads to reliance on limited, potentially biased datasets that may result in poorly generalizable results and biased disease outcome predictors.[4]

Several remediation strategies have been proposed, including improving data collection, conducting complete case analysis, modeling missingness in analyses, supplementing with additional data, and employing imputation methodologies.[5] Existing imputation methods, many of which leverage census data based on geolocation and correlations between people's surnames and their self-reported race and ethnicity,[6,7] achieve moderate accuracy and require access to protected health information (PHI), limiting their applicability.[8,9]

Molecular tumor profiling, an assay used in support of therapy decisions in cancer patients, is often accompanied by a wealth of multimodal RWD that, once de-identified, can be harnessed for research.[10] This can include clinical metadata, imaging, and molecular data, such as DNA variants on a set of cancer related genes and transcript sequences from different patient tissues.[11]

Inferring genetic ancestry, or more accurately, genetic similarity to reference populations,[12] from molecular testing sequencing data, offers a potential solution to the challenge of missingness in race and ethnicity data. The granularity of such inferences is contingent on the availability of allele frequency data across samples from reference populations, with the most common level of genetic

[*] Joint first authorship.

[†] Correspondence: francisco.delavega@tempus.com

ancestry inference being at super-population level categories, as described by the 1000 Genomes Project.[13] Although genetic ancestry is not equivalent to race or ethnicity, a strong correlation between these two concepts has been observed among US populations.[14,15] We propose to leverage this correlation and the genetic information available in molecular testing RWD using two methods — one heuristic and the other based on machine learning — to impute mutually exclusive race and ethnicity categories from genetic ancestry. Here, we benchmark these methods and find they outperform previously reported race and ethnicity imputation methods, with a machine learning-based method providing the most accurate imputation.

2. Methods

The categorizations of race and ethnicity in this study adhere to the standards developed by the US Office of Management and Budget,[16] which are also used in the US census. These standards are based on two self-reported questions: a) Race (American Indian or Alaska Native, Asian, Black or African American, Native Hawaiian or Other Pacific Islander, and White); and b) Ethnicity (Hispanic or Latino and Not Hispanic or Latino). However, these categories present analytical challenges due to the orthogonal race and ethnicity questions, and it is often more practical to consolidate answers to these two questions into non-overlapping classes,[17] defined in this study as: Hispanic or Latino, non-Hispanic (NH) Asian, NH Black, and NH White, with the other races having insufficient numbers at the moment to develop reliable models in our source data. This consolidation allows for a more streamlined and comprehensive analysis of race and ethnicity in the context of RWD.

2.1. *Data*

Genomic and clinical data from patients of multiple cancer diagnoses was obtained from the Tempus database. The selected cohort consisted of 132,523 de-identified records of patients whose tissues were sequenced with the Tempus xT next-generation sequencing (NGS) panel (596-648 genes, v2-v4, tumor-normal matched when tissue available)[11,18] from 2018 to 2022. These records had been previously de-identified for other studies and passed minimal data quality filters. A total of 33,232 records had populated race, ethnicity, and geolocation data and belonged to one of the four non-overlapping race and ethnicity categories that we imputed: 4,357 Hispanic or Latino, 1,258 NH Asian, 3,120 NH Black, and 24,497 NH White. Race and ethnicity information in the Tempus database is obtained from a combination of electronic health record integrations and data abstraction from clinical documents and can be self-declared by patients or observed by practitioners. Information could be missing because there was no attempt to collect it, because patients or practitioners abstained from answering, or because it was not captured in the Tempus database. Analyses were performed using de-identified data under human subject research exemption granted by Advarra, Inc. Institutional Review Board, protocol Pro00042950.

2.2. *Determination of genetic ancestry*

We estimated genetic ancestry proportions using a re-implementation of the ADMIXTURE supervised global genetic ancestry estimation algorithm.[19] This approach calculated the proportions of ancestries for five super-populations—Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS)—using a previously published bespoke set of 654 ancestry informative markers (AIMs).[20] Briefly, AIMs were selected from single-nucleotide variants present in the reference samples that intersect with the targeted regions of the Tempus xT NGS assay, are not protein-changing, and are present at significantly different frequencies across the reference populations.[21] We sourced reference allele frequency data for these AIMs from the 1,000 Genomes Project,[13] the Human Genome Diversity Project,[22] and the Simons Genome Diversity Project databases.[23] In the case of the AMR super-population, we excluded the 1,000 Genomes Project's admixed "AMR" population and only included allele frequencies for Native American individuals available in the other sources. To evaluate the accuracy of our methods, we compared our global ancestry proportion estimates on whole-genome sequencing data from the Pan-Cancer Analysis of Whole Genomes Project,[24] with published global ancestry proportions determined by summing genome-wide local ancestry segments derived using the RFMix method.[25] This comparison yielded an average mean squared error, normalized to the sum of population proportions present in the dataset, of 0.12. The Tempus xT assay utilizes matched normal tissue when available (present for 51% of the study cohort) to classify variants as either germline or somatic, but germline variants can still be inferred in the absence of normal tissue.[11] The genetic ancestry proportion estimation method utilizes variant calls from normal tissue or those deemed to be germline. To assess performance when no matched normal tissue is available, we estimated proportions from both the tumor sample and the matched normal sample for a subset of patients (N = 3,358) and found that the five estimated proportions were highly concordant, with Pearson's correlation coefficient ranging from 0.9977 to 0.9999.[20]

2.3. *Benchmarking and performance metrics for race and ethnicity category imputation*

We relied on our cohort's stated race and ethnicity data as available in the Tempus database as our ground truth. To assess the performance of imputation methods, we employed a range of accuracy measures specific to each predicted race or ethnicity category. *Recall*, also called *sensitivity* or *true positive rate,*[26] measures the proportion of individuals correctly assigned to a category among all individuals truly in that category. *Precision*, or *positive predictive value*,[26] is the fraction of relevant instances among the retrieved instances, i.e., the proportion of correctly assigned individuals among all those assigned to a category. The *F1-score* is the harmonic mean of precision and recall, providing a balance between these two metrics. We also evaluated several measures of overall accuracy. *Cohen's kappa*[27] is a measure of agreement between predicted and true categories, accounting for the possibility of agreement occurring by chance. The *correct rate*, or *accuracy,* measures the proportion of all predictions that are correct.[26] *Log loss* quantifies the difference between predicted probabilities of belonging to a class and the true value (0 or 1) of belonging to that class, with lower log loss indicating better model performance[28]. The area under the receiver operating characteristic curve, or *AUC*, is a measure of model performance based on sensitivity and specificity across all classification thresholds and thus is not sensitive to any specific chosen threshold. *prAUC* is an analogous measure based on precision and recall. A predicted probability threshold of >0.5 was used for all metrics that rely on a single classification for each subject.

In addition to these common measures, we also utilized metrics proposed by Elliot *et al.*[6] The *weighted error* compares the true prevalence of race/ethnicity in the validation dataset to the predicted prevalence, providing an indication of the overall error rate. The *weighted correlation* measures the weighted average correlation (calculated using vectors of indicators) between true race and ethnicity and imputed category for each of the four categories, with weights equal to true prevalence. Together, these metrics offer a comprehensive evaluation of the performance of our imputation methods.

2.4. *Heuristic imputation of race and ethnicity*

We initially imputed mutually exclusive race and ethnicity categories from genetic ancestry proportions using a set of heuristics (Table 1) in part derived from admixture proportions reported in the literature for Black and Hispanic or Latino groups in the United States.[15] We defined four categories: Hispanic or Latino, NH Asian, NH Black, and NH White. Patients who did not fit the categories defined by these heuristics were labeled "complex." This latter category could be considered a no-call, as patients classified as such are typically excluded from any downstream analyses, and for comparison with other methods described below.

Table 1. Race and ethnicity imputation heuristics from genetic ancestry. Super-population codes: AFR, Africa; AMR, Americas; EAS, East Asia; EUR, Europe; SAS, South Asia.

Imputed category	Super-population genetic ancestry thresholds
Hispanic or Latino	>10% AMR and >70% combined AMR, EUR, and AFR
NH Asian	>70% combined EAS and SAS
NH Black	>20% AFR, <10% AMR, and >70% combined AFR and EUR
NH White	>80% EUR and <10% AMR
Complex	Remaining patients not meeting above thresholds

2.5. *Machine learning imputation of race and ethnicity*

We also developed machine learning (ML)-based imputation methods, wherein an ML algorithm is trained to classify subjects into race and ethnicity categories based on genetic ancestry and other inputs. For all models, a single train+test and validation set was assembled from the 33,232 patient records with stated race and ethnicity that fit our imputation categories and with available home address 3-digit ZIP code. Features used by these models included genetic ancestry proportions for AFR, AMR, EAS, EUR, and SAS; US census division of patient's home state (nine geographic groupings of states defined by the US Census Bureau: Pacific, Mountain, West North Central, West South Central, East North Central, East South Central, South Atlantic, Middle Atlantic, and New England); and "demographic proportions," i.e., proportions of Hispanic or Latino, NH Asian, NH Black, and NH White residing in each patient's three-digit ZIP code tabulation area (ZCTA), as available from the 2021 5-year American Community Survey and mapped to three-digit ZIP codes using UDS Mapper.[29] We split the train+test and validation sets 90/10 while maintaining the US census division proportions in each set to ensure that the sets were aligned well for populations whose genetic ancestry proportions vary by U.S. geography, e.g., Hispanic or Latino.[15] We evaluated models using three groups of features: 1) *ML-ancestry*: genetic ancestry proportions only; 2) *ML-ancestry+geolocation*: genetic ancestry proportions and US census divisions; 3) *ML-ancestry+demographic*s: genetic ancestry proportions and demographic proportions.

We implemented all machine learning models in R using the caret package (v 6.0.94).[30] A number of models based on supervised training algorithms were evaluated, including models based on the random forest (method="rf") and gradient boosting (method="gbm") algorithms. We ultimately chose a boosted logistic regression algorithm (method="LogitBoost",[28] presented here) as it provided the ability to make no-call assignments and applied a probabilistic threshold in classification. Boosted logistic regression is a supervised machine learning algorithm that utilizes negative log-likelihood as a cost function. It iteratively builds decision trees to classify subjects, where each iteration is trained on a sample (with replacement) of the data in which subjects who were incorrectly classified in the previous round are more frequently sampled. The final classifier consists of a weighted combination of decision trees, where trees with lower log loss have more weight, and it returns the probabilities of belonging to each category for each subject. We chose to assign "No Call" to any subject with all probabilities ≤0.5. All models were trained using 10-fold cross validation. Grid expansion was performed to evaluate boosting iterations from 1 to 100 in intervals of 10. The optimal number of iterations and the final model were selected based on the lowest log loss value.

3. Results

3.1. *Comparison of performance of race and ethnicity imputation methods*

Table 2 summarizes the overall performance of the heuristic assignment method and each of the ML models. The ML model that utilized combined genetic ancestry proportions and demographic proportions (the proportions of the population in a patient's three-digit ZCTA belonging to Hispanic or Latino, NH Asian, NH Black, and NH White) achieved the best mean F1-score (0.957), Cohen's kappa (0.936), correct rate (0.974), log loss (0.122), AUC (0.982), and prAUC (0.946) whereas the heuristic method performed the worst by most metrics: mean F1-score 0.939, Cohen's kappa 0.903, correct rate 0.959, weighted correlation 0.876, and weighted error 0.009. The ML model that solely considered genetic ancestry proportions achieved the best weighted correlation (0.930) and weighted error (0.007), whereas the ML model that included geolocation in the form of the US Census district of a patient's home address state had intermediate performance by most metrics.

Table 2. Overall performance of race and ethnicity imputation methods for the validation set (N=3,319). Metrics that rely on a single classification threshold used a predicted probability of ≥0.5 for computation. Refer to section **2.5** for ML method descriptions. Best performing metric indicated with bold.

Imputation Method	Mean F1-Score	Cohen's Kappa	Correct Rate	Weighted Correlation	Weighted Error	Log Loss	AUC	prAUC
Heuristic	0.939	0.903	0.959	0.876	0.009	-	-	-
ML-ancestry	0.954	0.934	0.973	**0.930**	**0.007**	0.127	0.980	0.930
ML-ancestry+geolocation	0.955	0.935	0.973	0.926	0.009	0.131	0.979	0.898
ML-ancestry+demographics	**0.957**	**0.936**	**0.974**	0.928	0.013	**0.122**	**0.982**	**0.946**

When evaluating performance by category, we found that recall, precision, and F1-score were all at or above 0.932 for the NH Asian, NH Black and NH White categories (Table 3). Performance of all imputation methods was worst for the Hispanic or Latino category, with recall ranging from 0.859-0.887, precision from 0.833-0.964, and F1-score from 0.859-0.909.

Table 3. Performance of race and ethnicity imputation methods on validation set (N=3,319) per classification category. Refer to section **2.5** for ML method descriptions. Best performing metric for each category indicated with bold.

		Classification Category, N			
Metric	Imputation Method	Hispanic or Latino, 435	NH Asian, 130	NH Black, 301	NH White, 2,463
Recall	Heuristic	**0.887**	**0.983**	0.983	0.966
	ML-ancestry	0.876	0.962	0.993	0.987
	ML-ancestry+geolocation	0.877	0.969	0.983	0.988
	ML-ancestry+demographics	0.859	0.976	**0.993**	**0.990**
Precision	Heuristic	0.833	**0.935**	0.942	**0.985**
	ML-ancestry	0.938	0.933	0.967	0.981
	ML-ancestry+geolocation	0.941	0.932	**0.969**	0.981
	ML-ancestry+demographics	**0.964**	0.932	0.968	0.978
F1-Score	Heuristic	0.859	**0.959**	0.962	0.976
	ML-ancestry	0.906	0.947	**0.980**	**0.984**
	ML-ancestry+geolocation	0.908	0.950	0.976	**0.984**
	ML-ancestry+demographics	**0.909**	0.954	**0.980**	**0.984**

3.2. *Performance of heuristic method*

Perhaps unsurprisingly, the heuristic method for assigning race and ethnicity categories based on genetic ancestry proportions alone underperformed by all measures as compared to the ML models (cf. Table 3). For the Hispanic or Latino category (the most difficult to predict using the selected features), the heuristic method did have the highest recall (0.887), but this was achieved at the cost of low precision (0.833), also reflected in this method obtaining the lowest F1-score (0.859) for that category. The heuristic method did achieve the highest recall, precision, and F1-score for the NH Asian category. Overall, although the heuristic method did not perform as well as the ML method, its performance was not far behind, achieving an overall correct classification rate of ~96% compared to ~97% for the ML models. The no-call rate (i.e., patients assigned to the "complex" category) was 2.5%.

3.3. *Performance of ML-ancestry boosted logistic regression model*

We found that the boosted logistic regression model that utilized only genetic ancestry proportions improved upon the heuristic method for all overall performance metrics, with an overall correct classification rate of 97.3%. It had lower recall (0.876) but higher precision (0.938) for the Hispanic or Latino category than the heuristic method. The model had a recall of 0.962-0.993 for the three non-Hispanic categories, indicating that it correctly identifies the vast majority of patients in those categories and is usually correct in its predictions, with precision ranging from 0.933-0.981. The no-call rate was very low at 0.7%.

3.4. *Performance of ML models including geolocation and demographics*

Adding geolocation or demographic composition obtained from patients' home address ZCTA areas to the genetic ancestry proportions (*ML-ancestry+geolocation* and *ML-ancestry+demographics*) slightly improved model performance according to most metrics, yielding a correct classification rate of 97.3% and 97.4%, respectively. The *ML-ancestry+demographics* model had the best overall performance by all metrics except the less commonly used weighted metrics, which emphasize performance according to the true prevalence of each race and ethnicity category in the validation dataset. Individual category performance metrics followed a similar pattern to that of the *ML-ancestry* model. Notably, the *ML-ancestry+geolocation* model had the best precision for the Hispanic or Latino category (0.964), which may be desirable for use cases where correct predictions of this category are valued over high recall. The no-call rate was 1.1% for *ML-ancestry+geolocation* and 1.0% for *ML-ancestry+demographics*.

3.5. *Reclassification of stated race and ethnicity categories by imputation*

We selected the *ML-ancestry* model for further characterization because of its minimal input needs by applying it to the entire labeled dataset, regardless of whether geolocation data was available (N=35,229). The resulting confusion matrix (Table 4) compares the imputed categories with the stated race and ethnicity from the Tempus database, including the rate of no-calls and the number and fraction of misclassified records for each stated category. The confusion matrix for the validation dataset mirrors this table in terms of percentages (data not shown).

Table 4. Confusion matrix comparing imputed race and ethnicity category to stated category for the *ML-ancestry* model on all labeled data, including records without geolocation information (N=35,229). Percentage of each stated category and numbers of patients (in parentheses) are indicated in each cell. Total percentage and number of misclassified patients for each stated category is given in the last row.

	Stated category			
Imputed category	Hispanic or Latino	NH Asian	NH Black	NH White
Hispanic or Latino	**82.3% (4,059)**	0.2% (2)	0.5% (18)	0.8% (195)
NH Asian	0.8% (39)	**96.5% (1,285)**	0.2% (6)	0.2% (57)
NH Black	1.8% (91)	0.1% (1)	**97.7% (3,231)**	0.2% (49)
NH White	11.4% (560)	2.5% (33)	0.6% (19)	**98.5% (25,266)**
No Call	3.7% (180)	0.8% (10)	1.0% (32)	0.4% (96)
Misclassified	14.0% (690)	2.7% (36)	1.3% (43)	1.2% (301)

Additionally, Figure 1 provides a visual representation of the allocation of patients from their stated race and ethnicity to their imputed categories through a flow diagram.

The confusion matrix further indicates that the Hispanic or Latino category experienced the highest rates of no-calls (3.7%) and misclassifications (14.0%), whereas the NH White category had the lowest (0.4% and 1.2%, respectively). The flow diagram in Figure 1 illustrates that most patients were assigned to their stated category, with the majority of misclassifications occurring between Hispanic or Latino and NH White categories. Nevertheless, the overall misclassification rate of this model was very low at 0.9%.

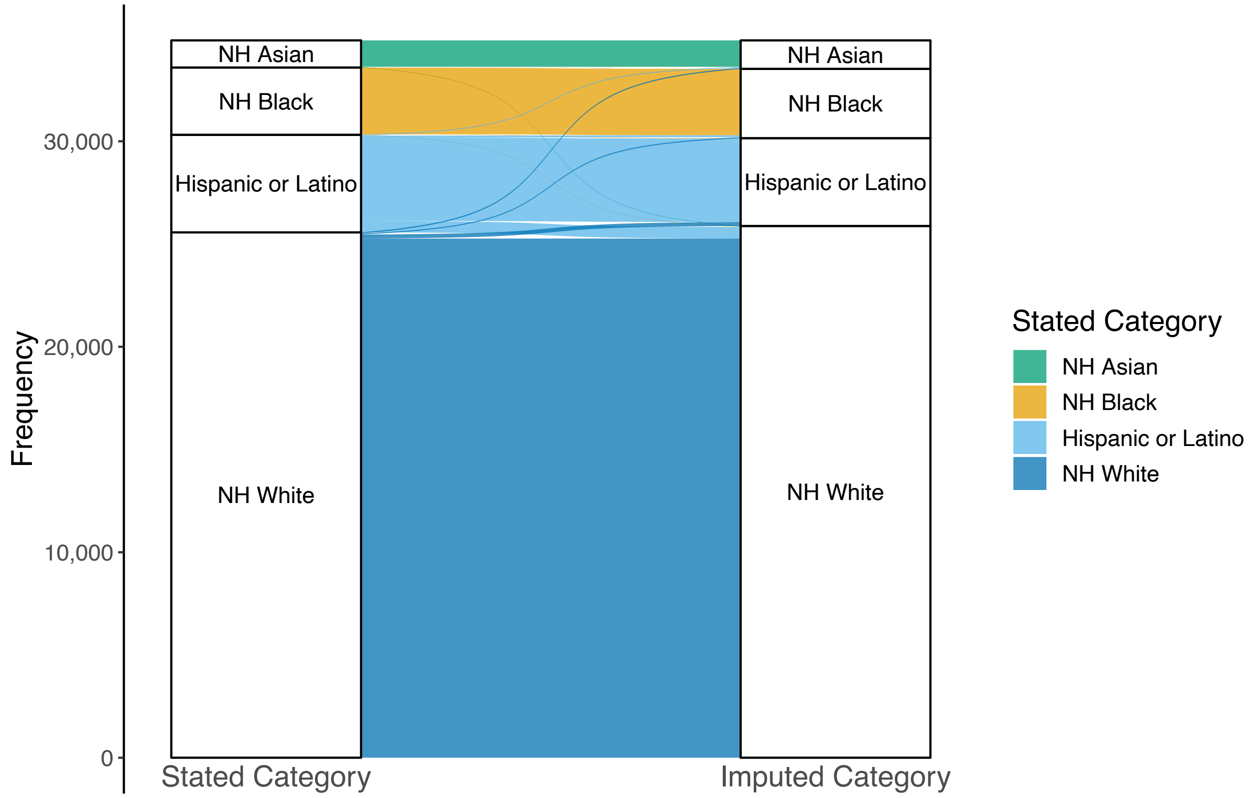

Fig. 1. Flow diagram showing the relationship between stated (left) and imputed (right) race and ethnicity categories with the *ML-ancestry* model in all labeled data, including records without geolocation information and excluding no-calls (N=34,911).

3.6. *Distribution of race and ethnicity categories imputed on unlabeled patients*

We also imputed race and ethnicity categories using the *ML-ancestry* model for all patients in the cohort (N=132,523) and examined the distribution of availability of race and ethnicity labels across categories (Figure 2). A total of 35,229 patients belonged to one of the four imputation categories according to their stated race and ethnicity data ("labeled"). There were 62,674 patients with no available race or ethnicity data at all ("unlabeled"), and an additional 34,620 with only partial information, i.e., either stated race or stated ethnicity (or both) were available, but patients did not fall into one of the four imputation categories, most frequently because ethnicity was unavailable (“partially labeled”). Imputed categories had comparable levels of unlabeled data, with the No Call and NH Asian categories having the most (53% and 52%, respectively) and NH Black having the least (44%). The Hispanic or Latino category had the highest level of labeled data by far (40%) due to the definition of that category only requiring a stated ethnicity of “Hispanic or Latino” and allowing stated race to be any value, including a missing value. The remaining categories had 22-26% labeled data. We observed that about half of each of the NH Asian, NH Black, and NH White imputed categories had records with a concordant stated race but a missing ethnicity (data not shown).

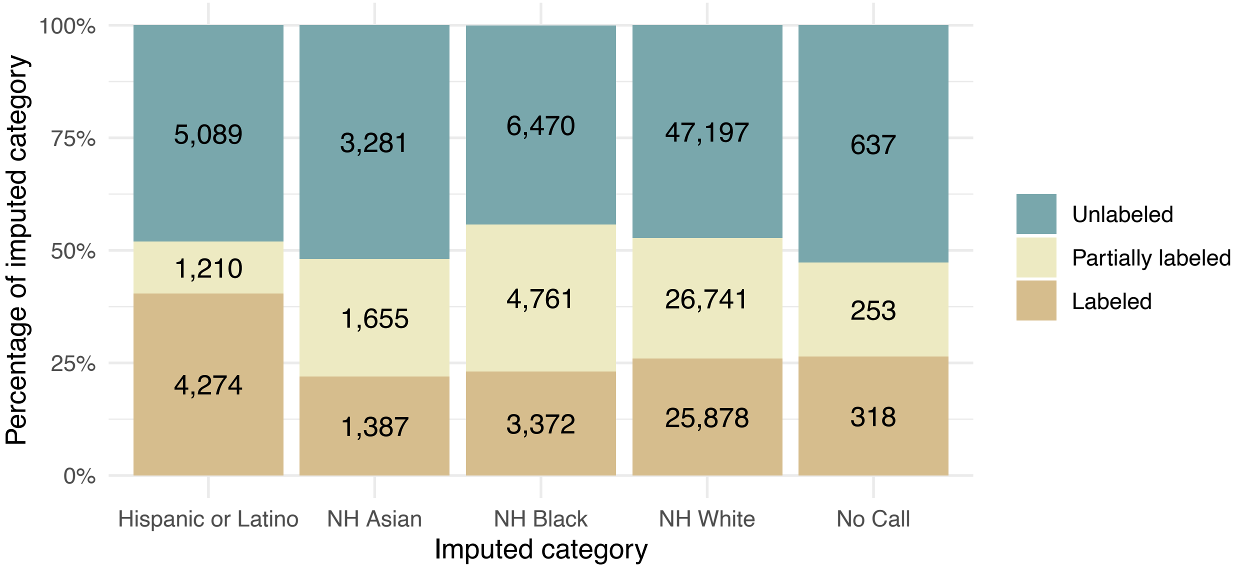

Fig. 2. Counts of patients in the full dataset (N=132,523) by label availability status and race and ethnicity category as imputed using the *ML-ancestry* model. Labeled = stated race and ethnicity are available, and a patient falls into one of: Hispanic or Latino, NH Asian, NH Black, or NH White based on this information. Unlabeled = neither stated race nor ethnicity is available. Partially labeled = either stated race or ethnicity is available, but the patient cannot be placed in one of the four listed categories.

3.7. *Analysis of potential biases*

The dataset used to develop our ML models is heavily imbalanced, with the largest group of patients (~74%) having a stated category of NH White, and the smallest group (~4%) having a stated category of NH Asian, potentially leading to overfitting to the majority category and biasing model performance. To address these potential problems, we evaluated additional models beyond those discussed here, wherein each model was trained in the same way except that each train+test set was downsampled to require an equal number of patients in each category, matching that of the category with the smallest number of patients. However, the downsampled models exhibited worse overall performance by all of our metrics and, within each category, had lower F1-scores (data not shown). Additionally, the performance metrics computed on the train+test sets during cross-validation were only slightly better than those computed on the validation set, alleviating concerns of overfitting. Importantly, the performance metrics we considered included metrics broken down by classification category to enable evaluation of whether any particular category was underperforming relative to the others. We also considered metrics that are suited to imbalanced data, such as Cohen's kappa.

4. Discussion

Although a direct comparison of our methods with other imputation methods was not possible due to the absence of PHI (such as surnames or addresses) in our de-identified dataset, we compared our performance to that reported in the literature. Our models consistently and substantially outperformed these prior methods.[6–9,31] E.g., the weighted correlation of our *ML-ancestry* model was 24-33 percentage points better, while its weighted error was an order of magnitude lower than other methods (Table 5).

Table 5. Comparison of performance metrics of *ML-ancestry* and other imputation methods based on metrics reported in the literature. Best performing metric indicated with bold. BISG = Bayesian Improved Surname and Geocoding;[6] CTBF = CT-based full;[9] CTBR = CT-based reduced.[9]

Imputation Method	Cohen's Kappa	Correct Rate	Weighted Correlation	Weighted Error	Reference
ML-ancestry	**0.934**	**0.973**	**0.930**	**0.007**	This study
BISG	0.58	0.78	0.597	0.089	Xue et al, 2019a[9]
CTBF	0.67	0.81	0.668	0.048	Xue et al, 2019a[9]
CTBR	0.65	0.81	0.595	0.051	Xue et al, 2019a[9]
Random Forest	0.67	0.807	0.672	0.025	Xue et al, 2019b[8]

In our study, the category with the lowest recall was Hispanic or Latino, ranging from 86-89%. This category also had the highest level of no-calls (3.4% vs. ≤1%). Prior methods report an even more pronounced drop in performance for this category.[6–9,31] However, the *ML-ancestry+demographics* model provided the best precision (96%) at a good recall rate (86%), while the Heuristic method provided the best recall (89%) but at a significantly lower level of precision (83%). Although the intended use of the imputation may dictate the best trade-off, we believe that precision is the most important feature as minimization of misclassified subjects is generally more desirable. The drop in performance in the Hispanic or Latino category may be due to the fact that self-affiliation with this category corresponds more with culture and language than with genetic similarity,[14] with levels of admixture within this group varying widely depending on country of origin and among the coasts of the US.[15]

As with all RWD analyses, our work has potential limitations. Differences between patients with complete vs. incomplete stated race and ethnicity could affect model training and therefore imputation performance. The unequal distribution of imputed categories in labeled and unlabeled data suggests that there are indeed some slight differences in the composition of patients who lack race and ethnicity data, with imputed NH Asian category most likely to be missing this information, but therefore also most able to benefit from imputation. Given the limited numbers of American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander individuals in our dataset as well as the insufficient public allele frequency information from these groups, we are unable to develop models to impute those categories, meaning those individuals will be misclassified, typically as Hispanic or Latino and NH Asian, respectively. As the Tempus database grows and additional AIM allele frequencies become available, our model could be retrained to enable classification using these additional categories. While the performance of our models on populations outside the US is unknown, or indeed with differently ascertained population samples, our results suggest that retraining with additional data pertaining to those populations could yield similar performance in other settings.

When developing our race imputation methods, we adhered to established recommendations for ethical imputation.[32] We audited input data for bias, scrutinized methodological choices for potential bias introduction, rigorously assessed the accuracy of the imputed data, and our aims are to use this data to study or reduce disparities. Our adherence to these guidelines underscores our commitment to the responsible use of race imputation in promoting equity in healthcare.

5. Conclusions

Addressing racial disparities is pivotal to advancing equity in precision medicine. However, the frequent unavailability of data disaggregated by race and ethnicity in RWD can lead to biased outcome predictors,[34] inadequate representation in clinical trials,[33] and poorly targeted policies, potentially exacerbating disparities.[34] While the ultimate goal is to have complete self-reported data for optimal race and ethnicity information, our study highlights the efficacy of using genetic ancestry data to impute these categories in a de-identified setting, mitigating the challenge of data sparsity for these data in RWD from US populations. Our approach could allow more accurate identification of racial disparities in certain healthcare settings where genetic data are available, contributing to the development of fair artificial intelligence predictors and more targeted and equitable healthcare interventions.

6. Acknowledgments

We appreciate the feedback received on this work from Funmi Olopade (University of Chicago), John Carpten (City of Hope), Jose Trevino (Virginia Commonwealth University), Carlos D. Bustamante (Stanford University), Joel Dudley, Calvin Chao, Ezra Cohen and Kate Sasser (Tempus Labs). We also acknowledge Frank Nothaft, Rafael Esleyer, Nick Riggan, and Arvind Prasad (Tempus Labs) for their invaluable assistance in procuring data needed for this work. We thank Vanessa Nepomuceno from the Tempus Publications team for copyediting the manuscript.

References

1. Dang, A. Real-World Evidence: A Primer. *Pharm. Med.* **37**, 25–36 (2023).

2. Resnic, F. S. & Matheny, M. E. Medical Devices in the Real World. *N. Engl. J. Med.* **378**, 595–597 (2018).

3. Studna, A. Executive Roundtable: The Rise of RWD in Clinical Research. *Applied Clinical Trials.* 28 September 2023, https://www.appliedclinicaltrialsonline.com/view/executive-roundtable-the-rise-of-rwd-in-clinical-research (2023).

4. Cullen, M. R. *et al.* A framework for setting enrollment goals to ensure participant diversity in sponsored clinical trials in the United States. *Contemp. Clin. Trials* **129**, 107184 (2023).

5. Cabreros, I., Agniel, D., Martino, S. C., Damberg, C. L. & Elliott, M. N. Predicting Race And Ethnicity To Ensure Equitable Algorithms For Health Care Decision Making. *Health Aff.* **41**, 1153–1159 (2022).

6. Elliott, M. N., Fremont, A., Morrison, P. A., Pantoja, P. & Lurie, N. A New Method for Estimating Race/Ethnicity and Associated Disparities Where Administrative Records Lack Self-Reported Race/Ethnicity. *Health Serv. Res.* **43**, 1722–1736 (2008).

7. Derose, S. F., Contreras, R., Coleman, K. J., Koebnick, C. & Jacobsen, S. J. Race and Ethnicity Data Quality and Imputation Using U.S. Census Data in an Integrated Health System. *Med Care Res Rev* **70**, 330–345 (2012).

8. Xue, Y., Harel, O. & Aseltine, R. Comparison of Imputation Methods for Race and Ethnic Information in Administrative Health Data. *2019 13th Int Conf Sampl Theory Appl Sampta* **00**, 1–4 (2019).

9. Xue, Y., Harel, O. & Aseltine, R. H. Imputing race and ethnic information in administrative health data. *Health Serv. Res.* **54**, 957–963 (2019).

10. Walther, Z. & Sklar, J. Molecular Tumor Profiling for Prediction of Response to Anticancer Therapies. *Cancer J.* **17**, 71–79 (2011).

11. Beaubier, N. *et al.* Integrated genomic profiling expands clinical options for patients with cancer. *Nat Biotechnol* **37**, 1351–1360 (2019).

12. National Academies of Sciences, Engineering, and Medicine, Committee on the Use of Race, Ethnicity, and Ancestry as Population Descriptors in Genomics Research. *Using Population Descriptors in Genetics and Genomics Research: A New Framework for an Evolving Field.* (National Academies Press (US), (2023).

13. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68 74 (2015).

14. Lu, C., Ahmed, R., Lamri, A. & Anand, S. S. Use of race, ethnicity, and ancestry data in health research. *Plos Global Public Health* **2**, e0001060 (2022).

15. Bryc, K., Durand, E. Y., Macpherson, J. M., Reich, D. & Mountain, J. L. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet.* **96**, 37 53 (2015).

16. Budget, O. of M. and. Standards for the classification of federal data on race and ethnicity. *Fed. Reg.* **62**, 58782 (1997).

17. Flanagin, A., Frey, T., Christiansen, S. L. & Committee, A. M. of S. Updated Guidance on the Reporting of Race and Ethnicity in Medical and Science Journals. *JAMA* **326**, 621–627 (2021).

18. Beaubier, N. *et al.* Clinical validation of the Tempus xO assay. *Oncotarget* **9**, 25826-25832 (2018).

19. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009).

20. Miyashita, M. *et al.* Molecular profiling of a real-world breast cancer cohort with genetically inferred ancestries reveals actionable tumor biology differences between European ancestry and African ancestry patient populations. *Breast Cancer Res.* **25**, 58 (2023).

21. Kosoy, R. *et al.* Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America. *Hum Mut.* **30**, 69–78 (2009).

22. Bergström, A. *et al.* Insights into human genetic variation and population history from 929 diverse genomes. *Science* **367**, 6484:eaay5012 (2020).

23. Mallick, S. *et al.* The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. *Nature* **538**, 201–206 (2016).

24. Campbell, P. J. *et al.* Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).

25. Maples, B. K., Gravel, S., Kenny, E. E. & Bustamante, C. D. RFMix: A Discriminative Modeling Approach for Rapid and Robust Local-Ancestry Inference. *Am J Hum Genet.* **8**, 93(2), 278-88 (2013).

26. Pepe, M. S. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford University Press (2004).

27. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **20**, 37–46 (1960).

28. Dettling, M. & Bühlmann, P. Boosting for tumor classification with gene expression data. *Bioinformatics* **19**, 1061–1069 (2003).

29. UDS Mapper: Zip code to ZCTA crosswalk. 28 September 2023, https://udsmapper.org/zip-code-to-zcta-crosswalk/.

30. caret: Classification and Regression Training. 28 September 2023, https://cran.r-project.org/web/packages/caret/index.html.

31. Grundmeier, R. W. *et al.* Imputing Missing Race/Ethnicity in Pediatric Electronic Health Records: Reducing Bias with Use of U.S. Census Location and Surname Data. *Health Serv Res* **50**, 946–960 (2015).

32. Brown, K. S., Ford, L., Ashley, S., Stern, A. & Narayanan, A. Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: Recommendations and Standards Guide. Urban Institute research report (2021).

33. Loree, J. M. *et al.* Disparity of Race Reporting and Representation in Clinical Trials Leading to Cancer Drug Approvals From 2008 to 2018. *JAMA Oncol* **5**, e191870 (2019).

34. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).