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[1]
[1] Tempus AI, Inc., Chicago, IL

INTRODUCTION

- 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.

METHODS

Cohort

- 132,523 de-identified records of patients whose tissues were sequenced with the Tempus xT NGS panel.
- 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.

Ancestry inference

We estimated genetic ancestry proportions for five super-populations—Africa (AFR), the Americas (AMR), East Asia (EAS), Europe (EUR), and South Asia (SAS)—using 654 ancestry informative markers and a re-implementation of the ADMIXTURE algorithm.[1]

Heuristic method

Mutually exclusive race and ethnicity categories were imputed using a set of heuristics derived in part from admixture proportions reported in the literature for Black and Hispanic or Latino groups in the United States[2]:

- 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

Machine learning methods

- Data were split into a train+test set, N= 29,909, and validation set, N=3,323.
- Models were trained using boosted logistic regression and three groups of features: 1) *ML-ancestry*: genetic ancestry proportions only; 2) *ML-ancestry+geolocation*: genetic ancestry and 9 US census divisions; 3) *ML-ancestry+demographics*: genetic ancestry and demographic proportions; i.e., the proportions of the population in a patient's 3-digit ZIP code belonging to Hispanic or Latino, NH Asian, NH Black, and NH White.

SUMMARY

- Race and ethnicity data are **frequently missing** in real world data
- Heuristic and machine learning methods can impute race and ethnicity from genetic ancestry with **high accuracy (>95%)**
- Machine learning imputation methods **outperform** heuristic methods

RESULTS

Table 1. Overall performance of race and ethnicity imputation methods for the validation set (N=3,319).

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**

Figure 1. Relationship between stated and imputed race and ethnicity

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

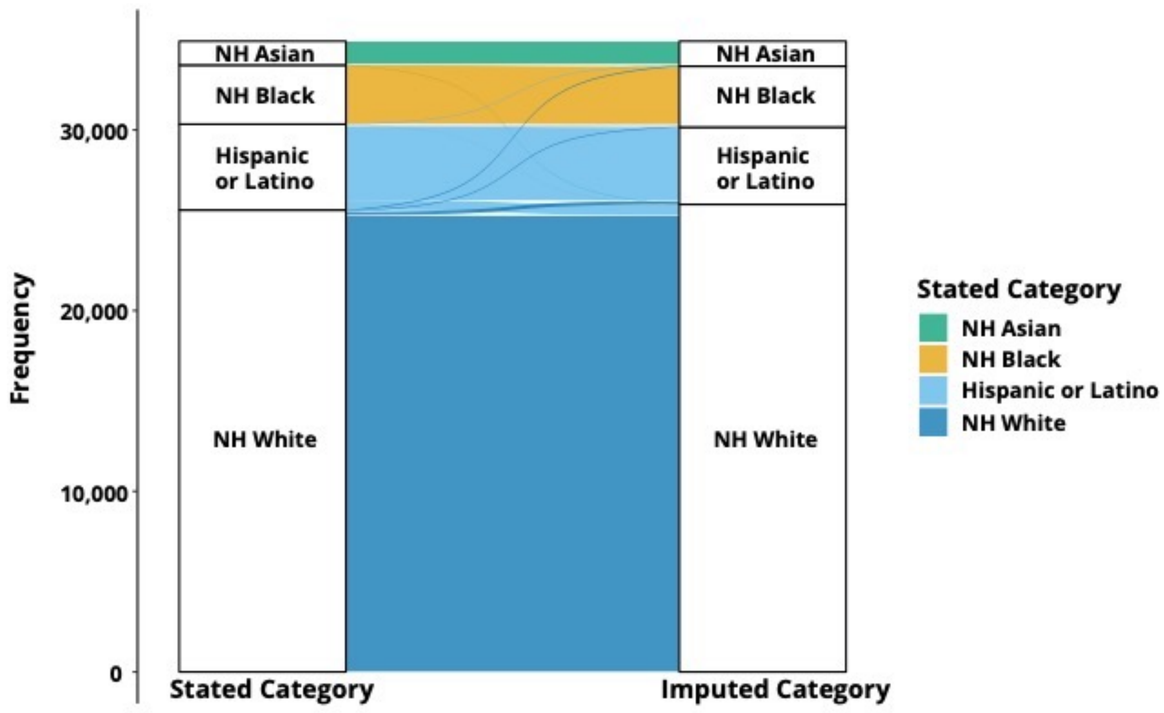

Figure 2. Race and ethnicity label availability status by imputed category

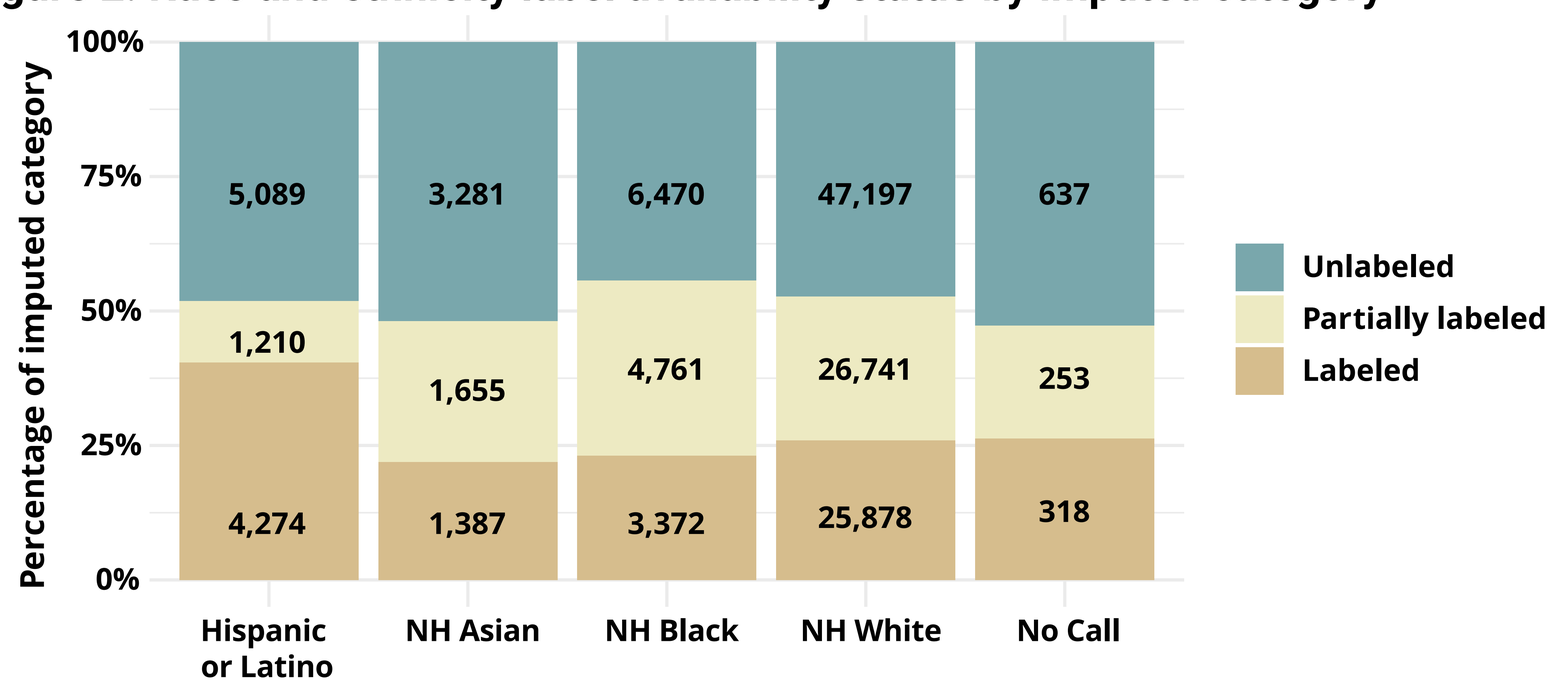

Fig. 2: Counts of patients in the full dataset (N=132,523) as imputed using the *ML-ancestry* model.

- 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.

References

1. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res* **19**, 1655-1664 (2009).
2. Bryc, K., Durand, E. Y., Macpherson, J. M., et al. The Genetic Ancestry of African Americans, Latinos, and European Americans across the United States. *Am J Hum Genet.* **96**, 37 53 (2015).

ACKNOWLEDGMENTS

We thank Vanessa M. Nepomuceno, Ph.D. from the Tempus Science Communications team for poster development.

Correspondence:
francisco.delavega@tempus.com