Augmenting Race and Ethnicity in a Real-World Oncology Cohort Using the Bayesian Improved Surname Geocoding (BISG) Methodology

Gene G. Ho, MPH[1]; Philani Mpofu, PhD[1]; Olive Mbah, PhD[1]; Cleo A. Ryals, PhD[1]

[1]Flatiron Health, New York, NY

MSR136

Scan to learn more

Background

- Race/ethnicity data missingness is a common challenge in real-world data (RWD) sources and a barrier to health equity and clinical trial diversity efforts
- In 2023, the Reagan-Udall Foundation for the FDA cited the Bayesian Improved Surname Geocoding (BISG) method as an approach for addressing race/ethnicity missingness in RWD
- BISG integrates an individual's name and census location to make predictions about their race/ethnicity; however, it's use has not been evaluated in the context of oncology RWD
- We examined the validity of BISG in an electronic health record (EHR)-derived cohort of US-based patients with cancer, assessing race/ethnicity concordance (EHR-documented vs BISG-imputed) and associations with patient outcomes

Methods

- **Data source:** The US-based Flatiron Health Research Database[1]
- **Setting:** The study included 2,250,391 patients diagnosed with cancer from January 1, 2011, to October 31, 2024
- **Statistical analysis:**
  - BISG[2] performance was assessed by examining accuracy, the area under (AU) the precision-recall curve (PRC), AU receiver-operator curve (ROC), and the kappa statistic
  - Patient characteristics of the EHR-documented and BISG-augmented cohorts were compared using standardized mean differences (SMDs)
  - Multivariable Cox models were applied to compare associations of EHR-documented versus BISG-imputed race/ethnicity with outcomes (ie, real-world overall survival [rwOS], time to treatment initiation, and clinical trial participation)

BISG is a valid and cost-effective approach for improving race/ethnicity completeness in oncology RWD, with implications for supporting regulatory and market access use cases related to assessing oncology drug safety, efficacy, and utilization among diverse patients in the postapproval setting

Future Directions

- Leveraging BISG and RWD to support monitoring of oncology drug safety, efficacy, and utilization among historically underrepresented patients in the postapproval setting
- Expanding BISG and the wru R package to unlock additional information for additional racial/ethnic groups (e.g., American Indian/Alaska Native; Native Hawaiian/Pacific Islander; or Middle Eastern/North African)
- Recalibrating BISG probabilities using additional data points (e.g., birth sex and age)

Acknowledgments: Darren Johnson (Flatiron Health) for publication management
Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. During the study period, GGH, PM, OM, and CR reported employment with Flatiron Health, Inc. and stock ownership in Roche. Data first presented at ISPOR 2025 (Montreal, QC, Canada; May 16, 2025)
Author contact information: Gene Ho, gene.ho@flatiron.com

References

1. Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed April 18, 2025. https://flatiron.com/database-characterization
2. Imai K, et al. *Political Anal*. 2016;24(2):263-272. doi:10.1093/pan/mpw001
3. Hand DJ, et al. *Mach Learn*. 2001;45(2):171-186. doi:10.1023/a:1010920819831

Results

- BISG implementation resulted in improved race/ethnicity capture for all patients, with the largest proportional increases observed among patients of color:
  - Latinx: +53.09%
  - NL-Asian: +38.14%
  - NL-Black: +34.69%
  - NL-White: +17.45%
- There were no patients with Unknown race/ethnicity after BISG implementation
- Overall, BISG performance was strong.
  - Overall AU PRC = 0.85
  - Overall AU ROC[3] = 0.95
  - Overall accuracy = 85.7%
  - Kappa statistic = 0.67
- Key patient characteristics were similar across EHR and BISG cohorts. All SMDs for selected important characteristics were below a prespecified threshold of 0.25; however, most were ≤ 0.05 (Figure 1)
- Outcomes analyses (ie, hazard ratios) of race/ethnicity were consistent across EHR-documented and BISG-imputed cohorts (Figure 2)

Figure 1. Standardized Mean Differences of Patient Characteristics in EDMs

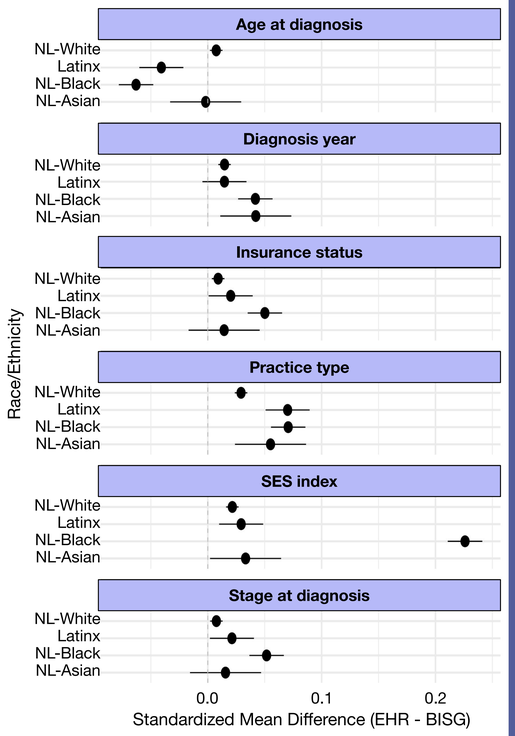
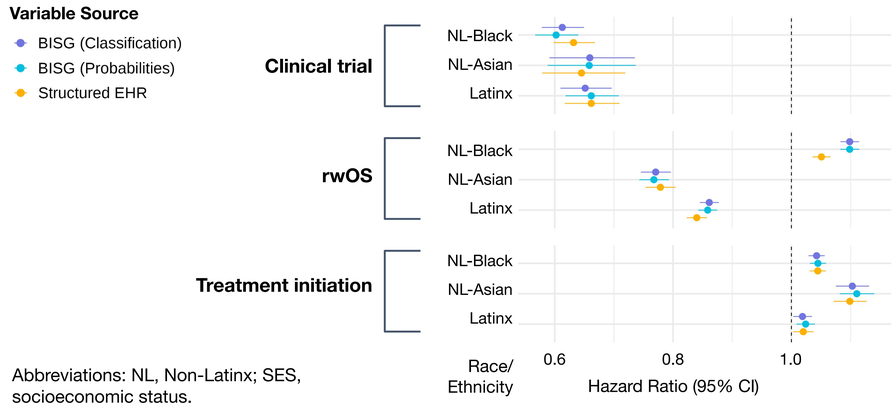


Figure 2. Comparison of Real-World Outcomes by Race/Ethnicity Data Source



Abbreviations: NL, Non-Latinx; SES, socioeconomic status.