Assessing Bias in LLM-Extracted Real-World Data: A Health Equity Analysis of Access to Care and Outcomes in Metastatic Breast Cancer

O. Mbah[1], G. Ho[1], C. Keane[1], Q. Yuan[1], C. Ryals[1]
[1]Flatiron Health, New York, NY

Background

- As large language models (LLMs) are increasingly used to extract clinical data from electronic health records, understanding their performance across subgroups (i.e., model fairness) is essential to mitigating bias in evidence generation
- We evaluated the reproducibility of scientific conclusions when using LLM-extracted vs. human-abstracted data to examine inequities in biomarker testing and overall survival (OS) among patients with HR+/HER2- metastatic breast cancer (mBC)

Methods

- **Data source:** This study used the US-based, electronic health record-derived deidentified Flatiron Health Research Database[1]
- **Setting:** The study included females ≥18 years diagnosed with HR+/HER2- mBC between January 2018 and January 2025. LLMs and human abstraction were used to derive clinical variables, including mBC diagnosis and dates, HR+/HER2- status, and biomarkers, from unstructured clinical documents
- **Main outcome measures:** Using both human abstracted and LLM-extracted variables, we: *(1) estimated the association between race/ethnicity, social determinants of health (SDOH), and biomarker testing (e.g., ESR1, PIK3CA)*, *(2) estimated Kaplan-Meier curves and median real-world overall survival (rwOS) by race/ethnicity and SDOH*, and *(3) modeled the association between race/ethnicity, SDOH and rwOS*
- **Statistical analysis:** We used multivariable Cox proportional hazard and Fine-Grey competing risk models to estimate hazard ratios (HRs) and 95% confidence intervals (CIs). To assess bias, we calculated the difference in effect estimates (e.g., HRs) and overlap of 95% CI between human abstracted and LLM-extracted cohorts

Results

- **Participants:** Patients in the LLM-extracted (N = 25,055) and human-abstracted cohorts (N=8530) mostly exhibited similar sociodemographic and clinical characteristics (Table 1)
- Across both cohorts, Latinx, Black, and Asian patients were generally less likely to undergo biomarker testing than White patients (e.g., Black vs. White, LLM: HR=0.89; 95%CI:0.84-0.93 vs. human: HR=0.91; 95%CI:0.83-0.99) (Figure 1)
- SDOH estimates were also similar across cohorts (e.g., patients from the most affluent neighborhoods were more likely to receive biomarker testing than patients from the least affluent neighborhoods [LLM: HR=1.23; 95%CI:1.17-1.29 vs. human: HR=1.28; 95%CI:1.17-1.40])
- rwOS estimates showed consistent survival differences by race/ethnicity and SDOH factors in both cohorts (Figures 2 & 3)

Table 1. Selected Demographic and Clinical Characteristics of Patients in Human-Abstracted and LLM Cohorts

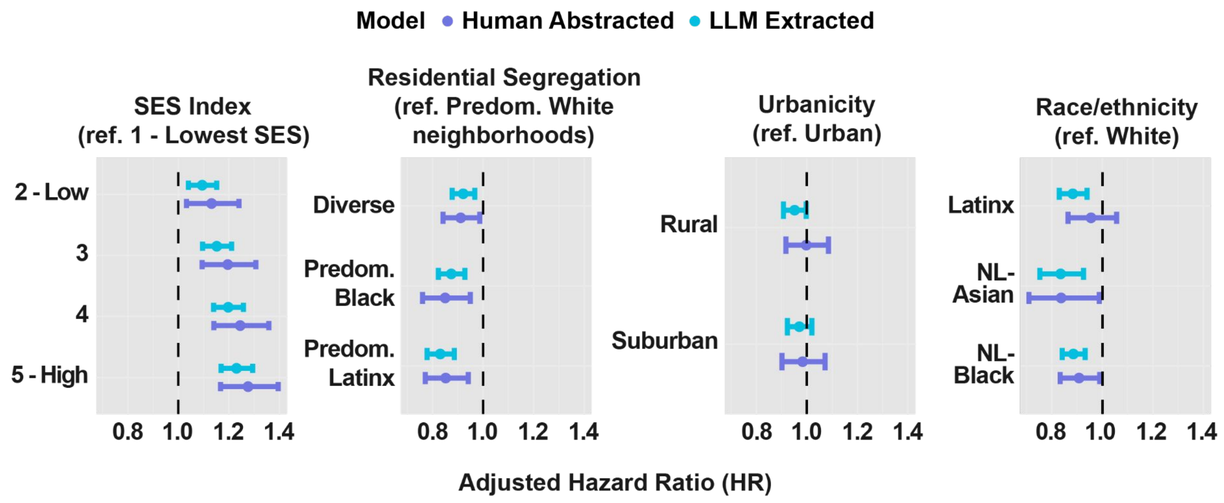
Characteristic	Human abstracted, No. (%) (n = 8530)	LLM-derived, No. (%) (n = 25,055)
Age, y		
18-49	1212 (14)	3342 (13)
50-64	2846 (33)	7972 (32)
65-74	2438 (29)	7266 (29)
75+	2034 (24)	6475 (26)
Race/ethnicity		
NL-Asian	224 (2.6)	536 (2.1)
NL-Black or African American	954 (11)	2573 (10)
Latinx	617 (7.2)	1683 (6.7)
NL-White	5282 (62)	14,828 (59)
Other Race/Missing	1453 (16.6)	5435 (22.1)
ECOG PS		
0	2749 (32)	8029 (32)
1	1933 (23)	5665 (23)
2-4	946 (11)	2750 (11)
Unknown/Not documented	2902 (34)	8611 (34)

NL: Non-Latinx; ECOG PS: Eastern Cooperative Oncology Group performance status. Percentages may not add up to 100% due to rounding

Acknowledgments: Darren Johnson, PhD (Flatiron Health, Inc.) provided publication support, and Madeline Morenberg and Marissa Jones (Flatiron Health, Inc.) provided design support. We acknowledge the Flatiron Health Equity Scientific Advisory Board for their support of this work. Data first presented at ISPOR EU in Glasgow, Scotland, UK on November 10, 2025.
Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. During the study period, OM, GH, CK, QY, and CR reported employment with Flatiron Health, Inc. and stock ownership in Roche.
Author contact information: Olive Mbah: olive.mbah@flatiron.com

Results (continued)

Figure 1. Association Between Race/Ethnicity and SDOH With Biomarker Testing: Comparing Human-Abstracted and LLM-Derived Cohorts



We observed concordance in results between the human-abstracted and LLM-derived cohorts, with overlapping HRs and 95% CI across key endpoints

Figure 2. Association Between Race/Ethnicity and SDOH With Overall Survival: Comparing Human-Abstracted and LLM-Derived Cohorts

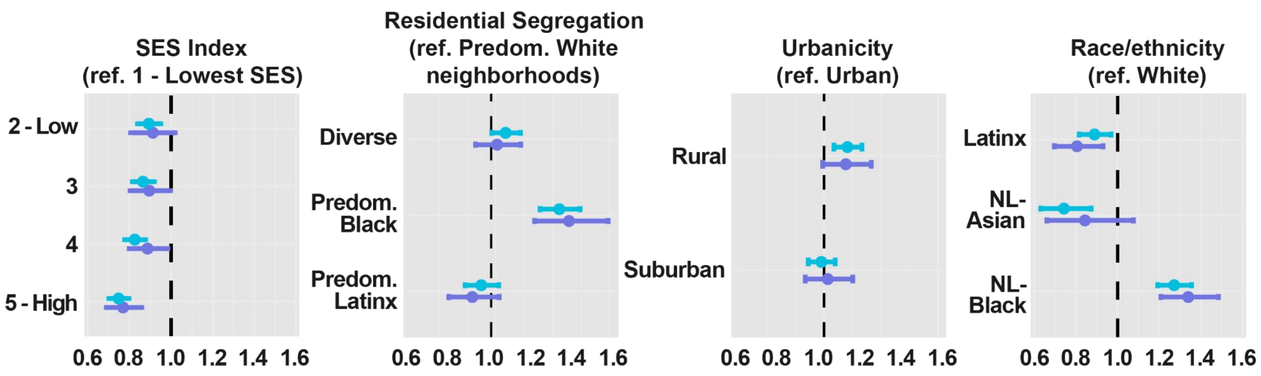
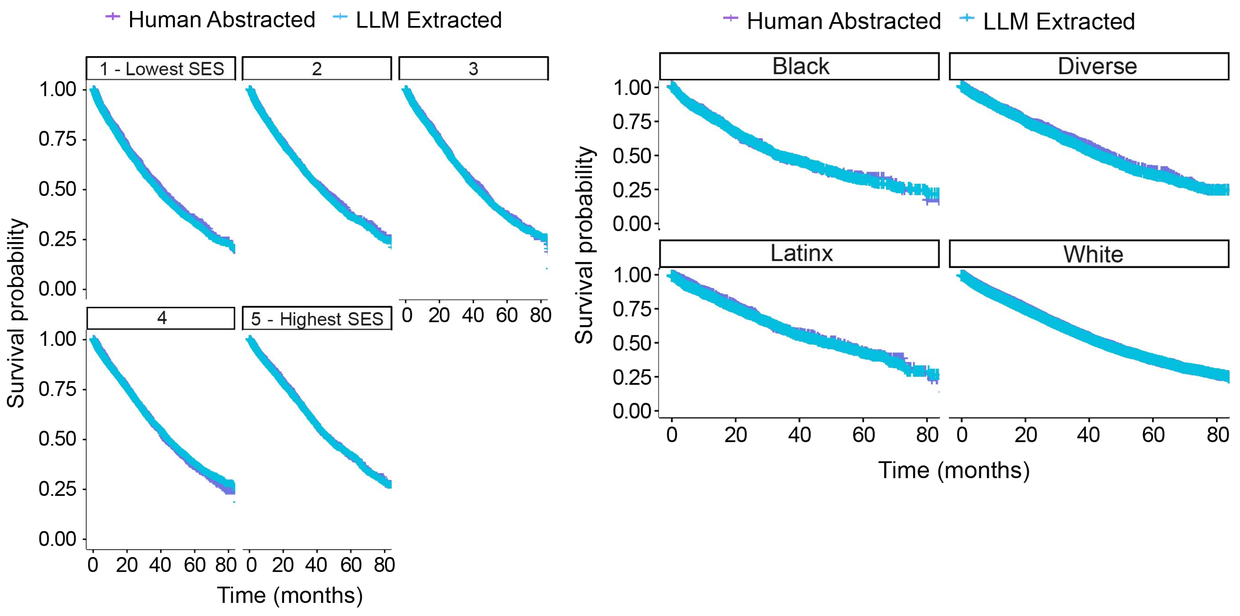


Figure 3. Real World Overall Survival: Comparing Human-Abstracted vs. LLM-Derived Cohorts



Highly consistent findings across **LLM-derived and human-abstracted data** provide **evidence of model fairness** and **support the use of LLM data in equity-focused** cancer research

Future Directions

- Study findings support the use of LLM-derived variables for equity-focused research and suggest that, with rigorous validation, LLM-derived data can serve as a scalable alternative to manual abstraction in oncology outcomes studies
- Results also underscore the need for continued strategies to address persistent racial/ethnic and community-level inequities in access to cancer care and survival, as novel therapies reshape cancer care

References
1. Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed September 25, 2025. https://flatiron.com/database-characterization

Scan for abstract and digital poster