Benchmarking Clinical Trial Diversity: Methods to Characterize Eligibility-Based Representativeness

Corey M. Benedum,[1] Sandra D. Griffith,[1] Selen Bozkurt,[2] Somnath Sarkar[1]

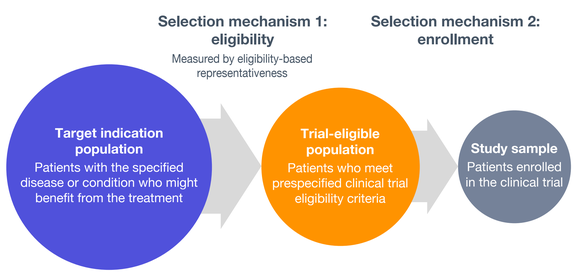
[1]Flatiron-Foundation Medicine Insights Lab, New York, NY, USA; [2]Department of Biomedical Informatics, Emory University, Atlanta, GA, USA

Poster HPR128

Background

- Clinical trials in oncology often fail to enroll patient populations that are representative of the broader population who may benefit from the treatment being investigated.[1,2]
- To ensure that new treatments provide clinical benefit to as many patients as possible, regulators and payers have highlighted the need for inclusive and diverse clinical trials.[3]
- Identifying and breaking down barriers to representativeness is challenging. In clinical trials, non-representativeness may result from restrictive eligibility criteria or difficulties in enrolling patients (**Figure 1**).
- To improve diversity, it is important to be able to benchmark the inclusivity of clinical trials by measuring the representativeness of the study populations.
- However, although several methods are available to measure representativeness, they have not been directly compared, and there are no recommendations regarding which methods to apply in real-world practice.
- Here we assessed the utility of three representativeness scoring methods to measure how clinical trial eligibility criteria may affect inclusion in trial-eligible populations.

Figure 1: Restrictive eligibility criteria and difficulties in enrollment may contribute to poor patient representativeness in clinical trials



Methods

- This study used the US nationwide Flatiron Health electronic health record-derived de-identified database. The Flatiron Health database is a longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction.[4,5]
- During the study period, the de-identified data originated from approximately 280 US cancer clinics (approximately 800 sites of care). The study included 50,263 patients diagnosed with advanced non-small cell lung cancer (aNSCLC) between January 1, 2011 and March 24, 2023.
- Target indication populations were constructed for each of nine phase 3 clinical trials of treatments for aNSCLC shown in **Table 1**.[6] These corresponded to the populations of patients in the database who may benefit from the treatment being investigated in each trial.

Table 1: Representativeness was assessed for nine clinical trials

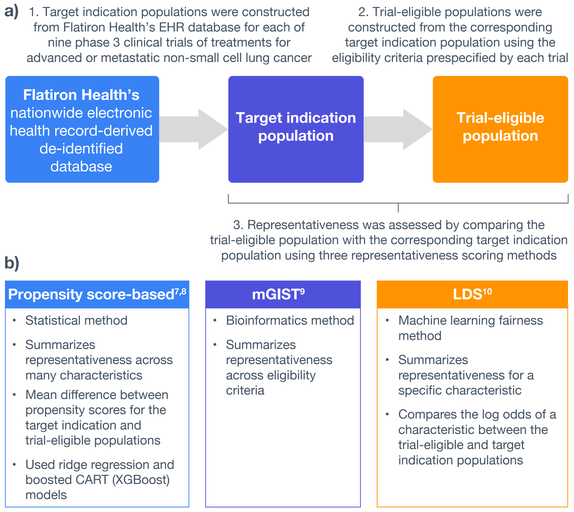
Trial name	Investigational treatment	Treatment line
BEYOND	Bevacizumab	1L
Checkmate017	Nivolumab	2L+
Checkmate057	Nivolumab	2L+
Checkmate078	Nivolumab	2L
Keynote010	Pembrolizumab	2L+
Keynote407	Pembrolizumab	1L
Keynote189	Pembrolizumab	1L
LUX8	Afatinib	2L
OAK	Atezolizumab	2L or 3L

1L, first line; 2L, second line; 2L+, second line or later; 3L, third line

Methods Continued

- From each target indication population, the trial-eligible population was constructed by applying the original prespecified eligibility criteria of the corresponding trial (**Figure 2a**).
- Focusing on selection mechanism 1 (**Figure 1**), the representativeness of patients in the trial-eligible population for each trial was assessed using three representativeness scoring methods (described in **Figure 2b** and **Table 2**). Representativeness metric scores were compared against literature-based benchmark values, if available.

Figure 2: Using real-world EHR data to a) construct target indication and trial-eligible populations, and b) assess the inequities in representativeness between target indication and trial-eligible patient populations



CART, classification and regression tree; EHR, electronic health record; LDS, log disparity score; mGIST, multidimensional generalizability index for study traits

Table 2: Comparison of metric inputs and outputs for propensity score-based, mGIST, and LDS methods for measuring representativeness in clinical trials

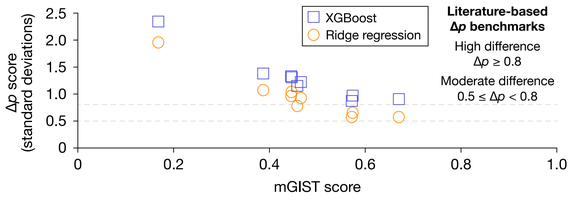
Metric inputs				
Features included in score	**Examples**	**Propensity score-based**	**mGIST**	**LDS**
Eligibility criteria features	• ECOG PS • Biomarker status • Vital signs and laboratory test values	Yes	Yes	No
Patient characteristic features	• Age • Race/ethnicity • Gender	Yes	No	Yes
Clinical and disease characteristic features	• Smoking status • Stage at diagnosis • Other vital signs and laboratory test values	Yes	No	No
Metric outputs				
		Propensity score-based	**mGIST**	**LDS**
Accounts for interdependencies between criteria/features		Yes	Yes	No
Representativeness granularity		Cohort level	Cohort level	Characteristic level
Derivation complexity		Medium	High	Low

ECOG PS, Eastern Cooperative Oncology Group performance status; LDS, log disparity score; mGIST, multidimensional generalizability index for study traits

Results

- Target indication populations ranged from 2,600 to 48,918 patients across the nine trials. Trial-eligible populations ranged from 1,638 to 21,364 patients.
- Mean differences ($\Delta\rho$) in propensity scores (in standard deviations) and multidimensional (overall) trial generalizability index for study traits (mGIST) scores are shown in **Figure 3**. For $\Delta\rho$, lower scores indicate greater similarity between the trial-eligible population and the target indication population (**Figure 4**). For mGIST, higher scores indicate greater similarity between the trial-eligible population and the target indication population.
- Representativeness scores calculated using ridge regression and XGBoost propensity score-based methods were positively correlated, and scores from both methods were inversely correlated with mGIST scores (**Figure 3**).
- Log disparity scores (LDSs) for selected patient demographic characteristics are shown in **Figure 5**.
- Clinical trial representativeness scores measured using each method are compared with literature-based benchmark measures in **Table 3**.

Figure 3: Correlation between propensity score-based and mGIST representativeness scoring methods

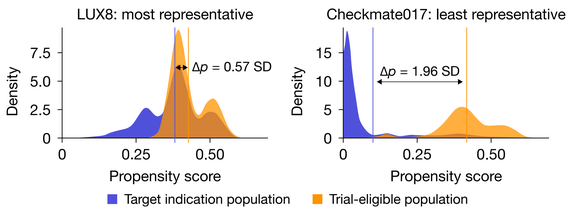


Correlation matrix: Pearson's *r*

	Δρ ridge regression	Δρ XGBoost	mGIST
Δρ ridge regression	–	–	–
Δρ XGBoost	0.997	–	–
mGIST	–0.956	–0.954	–

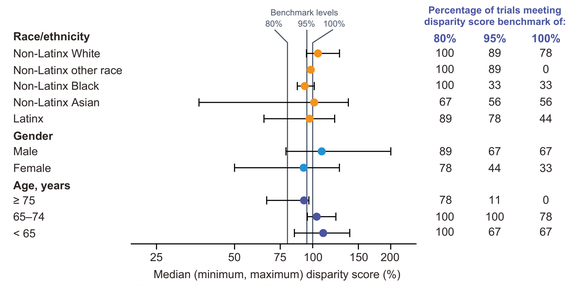
mGIST, multidimensional trial generalizability index for study traits; Δρ, mean difference

Figure 4: Differences in propensity score-based metrics for the most and least representative trials



Δρ, mean difference; SD, standard deviation

Figure 5: LDS for selected demographic characteristics



Disparity score = exp(LDS). Values below 100% indicate underrepresentation
LDS, log disparity score

Results Continued

Table 3: Comparison of clinical trial representativeness against benchmark values and recommendations for use of representativeness metrics in future trials

Comparison of representativeness versus literature-based benchmarks	
Propensity score-based	Using accepted literature benchmark values,[11] 4 out of 9 trials assessed using a propensity score-based (ridge regression) method had moderate representativeness and 5 had low representativeness
mGIST	Literature benchmark values for the mGIST method were not available
LDS	Most trials met the lower bound (≥ 80%) for representativeness proposed in the literature.[10,12] Fewer than 50% of trials met the 95% representativeness benchmark with respect to Black, female, or older (≥ 75 years) patients
Recommendations for use of representativeness scoring methods in future trials	
Propensity score-based	Case-specific benchmarks for 'good enough' have yet to be defined. We recommend evaluating a broad set of trials that includes approved and unapproved indications to empirically define meaningful benchmarks to guide clinical trial planners
mGIST	Additional research is required to identify meaningful benchmark values to describe trial representativeness
LDS	The lower bound of 80% is too low and overestimates the number of representative trials. We recommend that researchers aim for a higher benchmark (e.g. 95%) when evaluating representativeness for historically underrepresented populations

LDS, log disparity score; mGIST, multidimensional trial generalizability index for study traits

Summary and Conclusions

- Different methods can provide distinct but complementary insights to enable a comprehensive assessment of clinical trial representativeness. Combining methods to maximize representativeness may be important for projecting the most clinical benefit across diverse populations.

References

1. Fashoyin-Aje LA et al. *Clin Cancer Res*. 2023;29(18):3566–72.
2. Riner AN et al. *JNCI Cancer Spectr*. 2023;7(2):pkad009.
3. US Food and Drug Administration. Diversity Plans to Improve Enrollment of Participants from Underrepresented Racial and Ethnic Populations in Clinical Trials; Draft Guidance for Industry. 2022. Available from: www.fda.gov/regulatory-information/search-fda-guidance-documents/diversity-plans-improve-enrollment-participants-underrepresented-racial-and-ethnic-populations (Accessed March 21, 2024).
4. Ma X et al. Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR. medRxiv: 10.1101/2020.03.16.20037143 [Preprint; 86 pp.]. 2020. Available from: www.medrxiv.org/content/10.1101/2020.03.16.20037143v3 (Accessed March 21, 2024).
5. Birnbaum B et al. Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research. arXiv: 10.48550/arXiv.2001.09765 [Preprint; 25 pp.]. 2020. Available from: https://arxiv.org/abs/2001.09765 (Accessed March 21, 2024).
6. Liu R et al. *Nature*. 2021;592:629–33.
7. Stuart EA et al. *J R Stat Soc Ser A Stat Soc*. 2001;174(2):369–86.
8. Rogers JR et al. *J Biomed Inform*. 2021;119:103822.
9. Sen A et al. *J Biomed Inform*. 2016;63:325–36.
10. Qi M et al. *JAMIA Open*. 2021;4(3):ooab077.
11. Faranone SV. *PT*. 2008;33(12):700–11.
12. Equal Employment Opportunity Commission. Uniform Guidelines in Employee Selection Procedures. 1978. Available from: www.govinfo.gov/content/pkg/CFR-2011-title29-vol4/xml/CFR-2011-title29-vol4-part1607.xml (Accessed March 21, 2024).

Acknowledgments

This study was sponsored by Flatiron Health, Inc. Statistical programming and analytic support was provided by Christina Parrinello and Plinth Analytics. Medical writing support was provided by Adam Errington of PharmaGenesis Cardiff, Cardiff, UK, and was funded by Flatiron Health in accordance with Good Publication Practice (GPP 2022) guidelines (www.ismpp.org/gpp-2022).

Disclosures

CMB, SDG, and SS are employees of Flatiron Health, Inc. (an independent member of the Roche Group) and hold stock and/or equity ownership in Roche. SB has received consultancy fees from Flatiron Health, Inc.

Presented at the International Society for Pharmacoeconomics and Outcomes Research (ISPOR) 2024 meeting, May 5–8, 2024, Atlanta, GA, USA