Can ML-Extracted Variables Reproduce Real World Comparative Effectiveness Results From Expert-Abstracted Data? A Case Study in Metastatic Non-Small Cell Lung Cancer Treatment

Arjun Sondhi, PhD; Corey M. Benedum, MPH, PhD; Aaron B. Cohen, MD, MSCE; Sheila Nemeth, MPH, PhD; Selen Bozkurt, PhD

Flatiron Health Inc., New York, NY

Background

- Machine learning (ML) extraction of real world data (RWD) from unstructured text (e.g. clinical notes) in electronic health records (EHRs) is more cost-effective and scalable than manual human abstraction
- Standard ML model performance metrics only consider single variable accuracy in a vacuum; in practice, research is conducted using many variables for both cohort selection and statistical analysis
- Rigourous evaluation beyond standard ML metrics is needed to determine whether ML-extracted variables are fit for research use, including replication of use cases [1]
- In this work, we assess whether ML-extracted variables can replicate analytic conclusions obtained using expert-abstracted variables in an oncology comparative effectiveness analysis

Methods

- We compare real world overall survival (rwOS) of patients with metastatic non small cell lung cancer (mNSCLC) treated with first-line bevacizumab, carboplatin, and paclitaxel (BCP) to those treated with first-line carboplatin and paclitaxel (CP) [2]
- A sample of of 177,211 patients with a lung cancer ICD code was obtained from the nationwide (US-based) Flatiron Health longitudinal database, comprising de-identified patient-level structured and unstructured data, curated via technology-enabled abstraction [3, 4]
  - During the study period, the de-identified data originated from approximately 280 US cancer clinics (~800 sites of care)
- Two cohorts with a non-squamous mNSCLC diagnosis (2011-2022) who were Stage IV at diagnosis and receiving first-line treatment with BCP or CP were selected using either **expert-abstracted** or **ML-extracted** variables, along with additional structured variables (data cutoff 4/30/2022)
- We compared patient characteristics between these cohorts using standardized mean differences (SMD); an SMD > 0.1 was considered to indicate a meaningful difference
- After applying inverse propensity weighting to adjust for confounders, the hazard ratio (HR) of rwOS between treatment groups was estimated in both cohorts

Results

- After applying pre-specified inclusion-exclusion criteria, we obtain similar cohorts using abstracted as with ML-extracted data; the ML-extracted cohort achieved a precision of 0.84 and a recall of 0.80 with respect to the abstracted cohort
- These cohorts are broadly similar in baseline characteristics, with meaningful differences as indicated by SMDs > 0.1 observed for EGFR and KRAS mutation positivity (Table 1)
- The unadjusted median survival is slightly greater in BCP-treated patients compared to CP-treated patients; this is consistent in both cohorts (Figure 1)
- Inverse propensity weighting resulted in similar adjusted covariate balance in both the ML-extracted and abstracted cohorts (Figure 2)
- After weighting to adjust for confounders, the estimated HRs are (Figure 3):
  - **Abstracted cohort:** 0.8839 (95% CI [0.7376, 1.0593]) in favour of BCP
  - **ML-extracted cohort:** 0.8754 (95% CI [0.7297, 1.0502) in favour of BCP

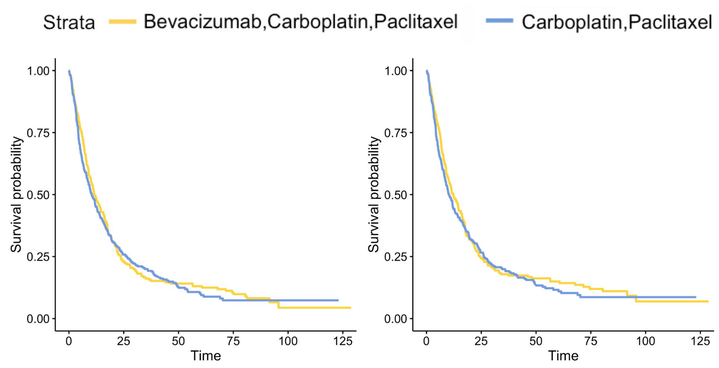

Figure 1. Unadjusted survival curves for rwOS within treatment groups. Left: abstracted cohort, right: ML-extracted cohort.

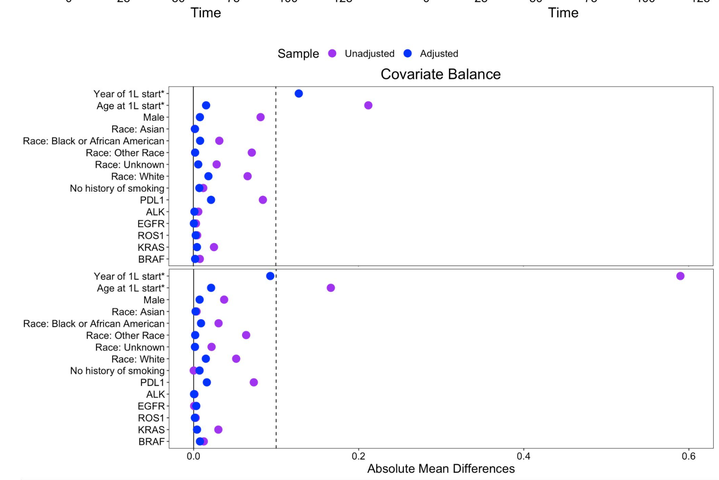

Figure 2. Covariate balance plots for adjusted confounders. Top: abstracted cohort, bottom: ML-extracted cohort.

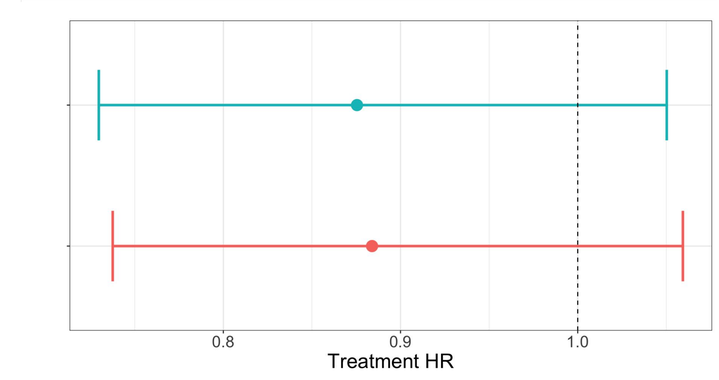

Figure 3. Estimated treatment hazard ratios for rwOS with 95% CIs. Top: abstracted cohort, bottom: ML-extracted cohort.

Table 1. Comparison of selected baseline characteristics of abstracted and ML-extracted cohorts. Additional characteristics compared were therapy start year, race, practice type, region, smoking status, and PD-L1/ROS1/BRAF biomarker status; all showed SMD < 0.1

	Abstracted cohort, N = 674	ML-extracted cohort, N = 643	SMD
Age at index (median, IQR)	66 (59, 73)	66 (59, 74)	0.02
Female sex	42%	44%	0.03
ECOG PS at index			0.01
0	34%	33%	
1	66%	67%	
KRAS mutation positive	16%	9.5%	0.18
ALK mutation positive	1.6%	0.8%	0.08
EGFR mutation positive	5.2%	2.5%	0.14

Conclusions

- Our study demonstrates that oncology RWD extracted using high-performing ML models may be used as an alternative to expert-abstraction
  - Our effect estimates from both cohorts also matched the direction of those observed in randomized clinical trials [2]
- ML-extracted RWD has the promise to unlock outcomes research at a massive scale, but proper evaluation is needed to assess research quality
- We conducted identical comparative effectiveness analyses using both ML-extracted and expert-abstracted variables, and showed minimal impact of ML model errors across multiple variables on analytic results
- Future research should continue to quantify the relationship between upstream ML model performance and downstream research results

References

1. Estevez, M, et al. "Considerations for the Use of Machine Learning Extracted Real-World Data to Support Evidence Generation: A Research-Centric Evaluation Framework." Cancers 14.13 (2022): 3063.
2. Sandler, A, et al. "Paclitaxel–carboplatin alone or with bevacizumab for non–small-cell lung cancer." New England Journal of Medicine 355.24 (2006): 2542-2550.
3. Birnbaum, B, et al. "Model-assisted cohort selection with bias analysis for generating large-scale cohorts from the EHR for oncology research." arXiv preprint arXiv:2001.09765 (2020).
4. Ma, X, et al. "Comparison of population characteristics in real-world clinical oncology databases in the US: Flatiron Health, SEER, and NPCR." Medrxiv (2020).

Disclosures

This study was sponsored by Flatiron Health, Inc., which is an independent subsidiary of the Roche Group. At the time of the study, all authors reported employment at Flatiron Health, Inc. and stock ownership in Roche.

Contact: Arjun Sondhi, PhD / arjun@flatiron.com