Performance Assessment and Validation of Real-World Response Data Generated Using a Deep Learning-Based Natural Language Processing Model Across Multiple Solid Tumors

MSR142

Kelly Magee[1], Qianyu Yuan[1], Auriane Blarre[1], Aaron B. Cohen[1], Aaron Dolor[1], Konstantin Krismer[1], Tori Williams[1], Qianyi Zhang[1]
[1]Flatiron Health, New York, NY

Background

- Response to treatment and related endpoints are critical in oncology clinical research. Rapid evidence generation in real-world cohorts can inform clinical trial study design and drug development
- We developed a real-world response (rwR) approach ("Scaled rwR") by leveraging natural language processing (NLP)-based deep learning models trained on expert human-abstracted data to generate rapid insights across large cohorts of patients
- This study aimed to describe the reliability, completeness, and internal validity of a novel machine learning (ML)-generated real-world response (rwR) approach applied to 14 of the most common solid tumor types

Methods

Data Source: The US-based, longitudinal Flatiron Health Research Database (FHRD), an electronic health record (EHR)-derived, deidentified database, comprising patient-level data originated from ~280 US cancer clinics (~800 sites of care; primarily community oncology settings) and curated via technology-enabled abstraction and artificial intelligence-based extraction methods, including NLP, ML, and large language models[1]

Setting:

- The study included a feasibility cohort for initial training and testing of the approach. The feasibility cohort was a convenience sample of multiple cohorts where human-abstracted data were available. These cohorts spanned multiple solid tumors, line settings, treatment types, and biomarkers
- Following initial training and testing with the feasibility cohort, the approach was evaluated in broader, solid tumor disease-based cohorts spanning the following cancer types: advanced urothelial (aUro), metastatic breast (mBC), metastatic colorectal (mCRC), advanced endometrial (aEndo), advanced gastric/esophageal (aGastric), hepatocellular (HCC), advanced head and neck (aHN), advanced melanoma (aMel), advanced non–small cell lung (aNSCLC), ovarian, metastatic pancreatic (mPanc), metastatic prostate (mPC), advanced renal cell carcinoma (aRCC), and small cell lung cancer (SCLC), including patients diagnosed between January 1, 2011, and July 31, 2024. The aNSCLC cohort was evaluated using two available datasets: one sampled (aNSCLC-1) and one broader unsampled dataset (aNSCLC-2) to assess performance across datasets

Variable: Scaled rwR was generated by a deep learning-based, NLP model, designed based on an existing human-abstracted rwR approach,[2] to extract clinicians' documentation of changes in disease burden (complete response [CR], partial response [PR], stable disease, progressive disease, or unknown) at imaging time points

Statistical Analysis:

- A subset of human-abstracted rwR data from the feasibility cohort was used to train the model. Reliability was established by assessing the correlation between real-world response rate (rwRR) using human-abstracted rwR data and Scaled rwR data in a test subset of the feasibility cohort
- Completeness was evaluated in the broad, disease-based cohorts by examining the proportion of treated patients with at least one assessment, as well as the time to first, second, and third assessments—reported as median, mean, and IQR for first (1L), second (2L), and third (3L) lines of therapy
- Internal validity was assessed by evaluating real-world overall survival (rwOS) for responders (those who ever achieved CR or PR) vs nonresponders (those who never achieved CR or PR) for the most frequent regimens in 1L to 3L for each disease (with ≥30 patients)

Results

- Within the test subset of the feasibility cohort (n = 4047), the correlation between human-abstracted and Scaled rwRR was $r = 0.86$. There was no meaningful variation when looked at by disease, line setting, or therapy class
- The broad solid tumor cohorts included 3406 to129 807 treated patients. Between 57.8% and 80.6% of patients had at least 1 assessment, with a median of first to third assessments within 1L, 2L, and 3L (See **Table 1** and supplementary material)
- Median times to first assessments for 1L to 3L ranged from 1.9 to 4.4 months. Median times to second, and third assessments ranged from 3.8 to 8.7 months, and 5.7 to12.9 months, respectively (See **Table 1** and supplementary material)
- Across the most frequent 1L to 3L regimens for each disease, responders for each cohort had significantly longer survival compared with nonresponders ($P < .05$) (See **Figure 1** and supplementary material)

Conclusion and Future Directions

- The availability of human-curated rwR data is foundational for developing effective ML-based data generation approaches
- The very strong correlation[3] between human-abstracted rwRR and ML-extracted rwRR generates confidence in the reliability of this new approach to data curation
- Assessments of data completeness were consistent with previous study[2] and clinical expectations. For example, longer times to first, second, and third assessments were observed in prostate cancer, where response assessment may be supplemented by tumor marker measurements and may not rely as heavily on imaging as compared with SCLC, where the shorter times to assessment are consistent with the aggressive nature of the disease
- The longer survival among ML-extracted rwR-defined responders vs nonresponders further supports the validity of the use of this variable
- Rapid advancements in ML technologies, particularly large language models, will likely allow for further refinement in capturing real-world response data

Table 1. Assessment Completeness and Frequency in First-Line

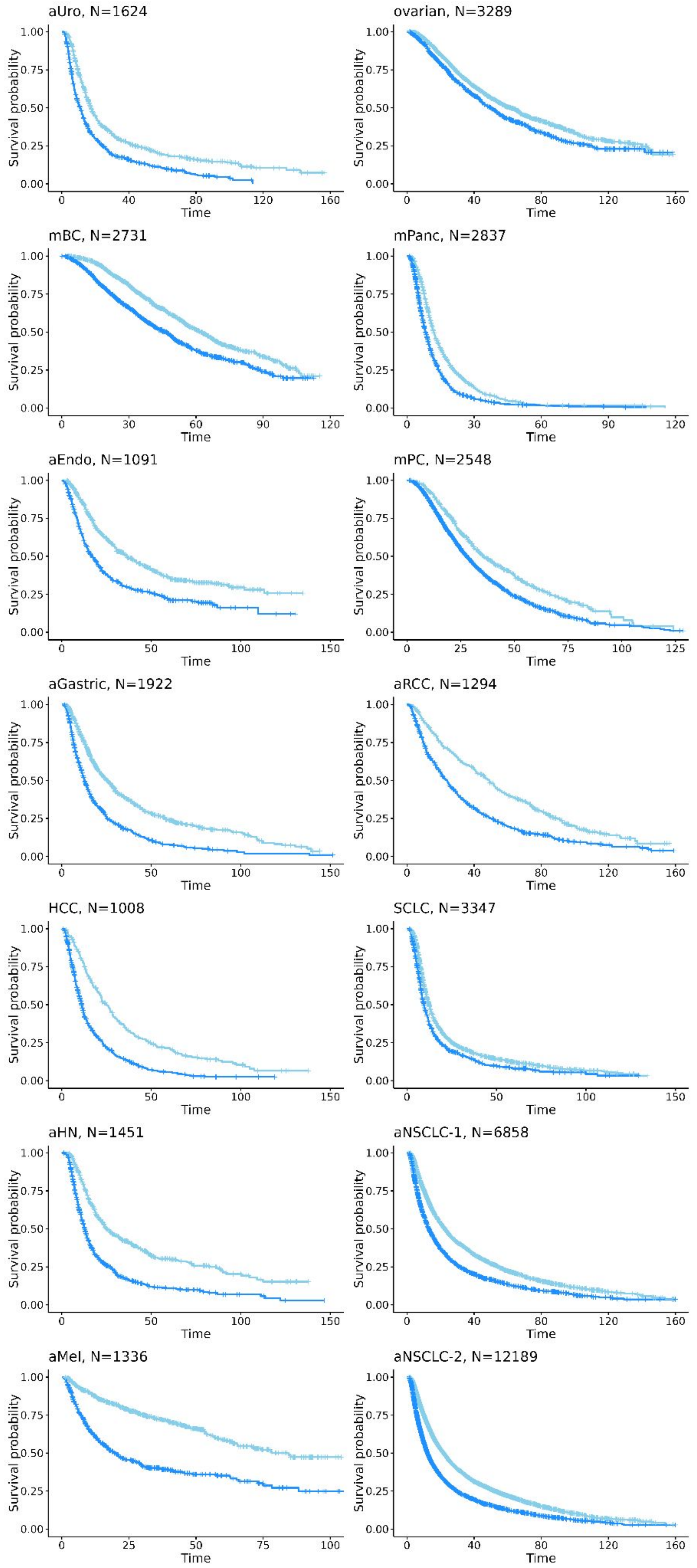
			Number of assessments		Time to first assessment, mo	
Disease	Pts, No.	Pts with ≥1 assessment, %	Mean	Median (IQR)	Mean	Median (IQR)
aUro	10,142	70.4	3.5	2 (1-4)	2.8	2.3 (1.7-3.1)
mBC	34,017	66.3	3.7	2 (1-5)	4.5	3.0 (2.0-4.5)
mCRC	32,848	74.3	3.9	3 (1-5)	3.1	2.5 (1.8-3.2)
aEndo	3406	69.2	3.2	2 (1-4)	3.6	2.4 (1.9-3.9)
aGastric	11,761	67.2	3.0	2 (1-4)	2.9	2.4 (1.8-3.1)
HCC	5053	53.8	3.0	2 (1-4)	3.2	2.5 (1.8-3.4)
aHN	9405	73.4	2.9	2 (1-3)	3.4	2.7 (1.9-4.0)
aMel	9889	70.8	5.3	3 (1-7)	3.6	2.6 (1.8-3.2)
aNSCLC-1	73,480	69.0	4.1	3 (1-5)	2.8	2.1 (1.6-2.9)
aNSCLC-2	129,807	69.4	4.0	3 (1-5)	2.7	2.2 (1.6-2.9)
Ovarian	8576	75.8	3.5	3 (2-4)	4.7	2.8 (1.9-4.6)
mPanc	12,232	59.1	2.8	2 (1-4)	2.5	2.2 (1.7-2.8)
mPC	12,835	62.5	2.3	2 (1-3)	7.0	4.4 (2.6-8.4)
aRCC	10,781	72.5	4.4	3 (1-6)	3.3	2.6 (1.9-3.3)
SCLC	9817	73.8	3.9	3 (2-5)	2.6	2.1 (1.4-2.9)

Abbreviations: IQR, interquartile range; Pts, Patients

Main Findings/Key Takeaways

This study **establishes the performance and validation** of a **novel ML approach** for **capturing rwR data from electronic health records**; supporting the **efficient and reliable generation of valuable outcome data across large cohorts**

Figure 1. rwOS for Responders vs Nonresponders in First-Line



Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. During the study period, KM, QY, AB, ABC, AD, KK, TW, and QZ reported employment with Flatiron Health, Inc. and stock ownership in Roche. Data first presented at ISPOR 2025 in Montreal, QC, Canada on May 16, 2025.

Author contact information: Kelly Magee, kmagee@flatiron.com

References

1. Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed April 21, 2025. https://flatiron.com/database-characterization
2. Ma X et al. *Adv Ther. 2021*. doi: 10.1007/s12325-021-01659-0
3. Swinscow TDV. 11. Correlation and regression. bmj.com. Accessed April 21, 2025. https://www.bmj.com/about-bmj/resources-readers/publications/statistics-square-one/11-correlation-and-regression

Scan for supplementary material