Assessing Quality of a Large Language Model (LLM)-Derived Prostate Cancer (PC) Real-World Dataset: An Application of the Validation of Accuracy for LLM/ML-Extracted Information and Data (VALID) Framework

MSR65

Scan to learn more

Patrick J. Ward, PhD, MPH[1]; Yunzhi Qian, PhD, MPH[1]; Eunice A. Hankinson, MSN, FNP-C[1]; Aaron Dolor, PhD[1]; Melissa Estevez, MS[1]

[1]Flatiron Health, New York, NY

Background

- The VALID Framework assesses LLM-derived real-world data (RWD) quality across three dimensions: variable-level metrics, verification checks, and replication analyses[1]
- This study applied VALID to a novel, LLM-derived PC dataset to determine suitability for generating real-world evidence (RWE)

Methods

- **Data source:** All cases of PC spanning non-metastatic through mCRPC setting from the US-based, electronic health record-derived, deidentified Flatiron Health Research Database.[2] LLM-derived data (n=385,566) were compared to human abstracted metastatic PC dataset (n=29,471)
- **Variables:** LLMs extracted clinically meaningful characteristics, including initial/metastatic diagnosis, castration-resistant PC (CRPC) or hormone-sensitive PC (HSPC) status, and treatment information. As a reference standard, the same variables were curated via expert human abstraction
- **Statistical methods:** The three dimensions of VALID were used as tools to assess LLM-derived data quality:
  - Variable-level metrics: The test set underwent double abstraction. F1 scores were calculated using Abstractor 2 as the reference standard for both Abstractor 1 and the LLM. The Delta F1 (LLM performance minus abstractor-to-abstractor agreement) was evaluated to assess whether the LLM achieved human-level performance
  - Verification checks: Verification checks assessed how often patients received >1 metastatic hormone-sensitive prostate cancer (mHSPC) line of therapy (LOT) in both the LLM-derived dataset and the expert human abstracted dataset
  - Replication analyses: Replication assessed real-world overall survival (rwOS) in two LOT defined cohorts—1L androgen receptor pathway inhibitor (ARPI) in metastatic CRPC (mCRPC) setting and 2L poly (ADP-ribose) polymerase inhibitors (PARPi) in mCRPC setting—in both the LLM-derived dataset and the human-abstracted dataset using the Kaplan-Meier (KM) method. Median rwOS was calculated and compared across datasets
- **Study cohorts:**
  - For Variable-level metrics, test sets of 349-500 patients were doubly abstracted to assess abstractor vs abstractor performance
  - For verification checks and replication analysis, an expert-human abstracted metastatic PC dataset was compared to an LLM-extracted prostate cancer dataset

Results

- Variable-level metrics:

Variable name	Delta F1 score (LLM – abstraction)
Initial diagnosis date	- 2.10%
Metastatic diagnosis date	- 2.11%
CRPC/HSPC status	- 0.52 %

- Verification checks:
  - In the LLM-derived dataset, the percentage of patients with >1 mHSPC LOT was 3.3% higher in the LLM-extracted dataset compared to the expert human-abstracted dataset
- Replication:
  - Median survival time (months, 95% CI) for patients with 1L mCRPC ARPI is similar in the LLM-extracted (25.3, 24.9-25.8) and expert human-abstracted (24.4, 23.7-25.1) datasets **(Figure 1)**
  - Median survival time (months, 95% CI) for patients with 2L mCRPC PARPi is similar in the LLM-extracted (15.8, 13.9-17.2) and expert human-abstracted (15.9, 14.5-17.8) datasets **(Figure 2)**

Figure 1. rwOS in 1L mCRPC ARPI-treated patients

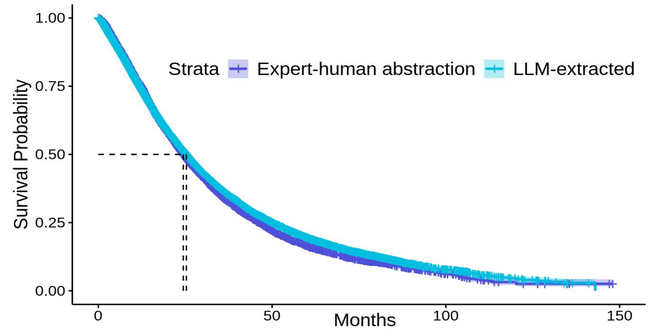
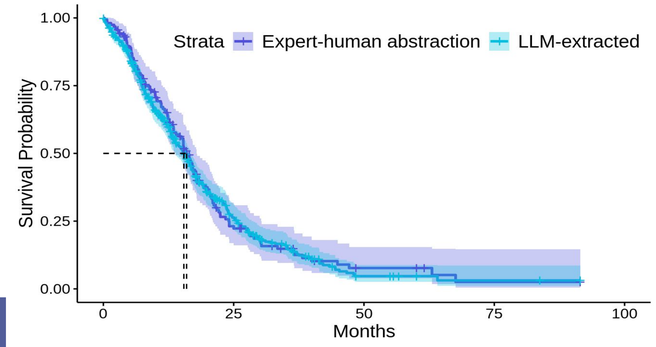


Figure 2. rwOS in 2L mCRPC PARP inhibitor-treated patients



The **VALID Framework** provided an approach that ensures **LLM-extracted RWD** can generate **fit-for-purpose RWE** in PC

Future Directions

- The VALID framework demonstrated fit-for-purpose performance across evaluated endpoints and populations. Future work will focus on scaling application and further standardizing evaluation across diverse real-world use cases

Disclosures: This study was sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. During the study period, PJW, YQ, EAH, AD, and ME reported employment with Flatiron Health, Inc. and stock ownership in Roche. Data first presented at ISPOR 2026 in Philadelphia, PA, USA on May 18, 2026.
Contact information: Yunzhi Qian, yunzhi.qian@flatiron.com

References

1. Estevez, M, et al. Ensuring Reliability of Curated Electronic Health Record–Derived Data: The Validation of Accuracy for Large Language Model–/Machine Learning–Extracted Information and Data (VALID) Framework. *JCO Clin Cancer Inform*. 2026; (10). doi:10.1200/cci-25-00215

2. Flatiron Health. Database Characterization Guide. Flatiron.com. Published March 18, 2025. Accessed April 9, 2026. https://flatiron.com/database-characterization