Extending the Validation of Accuracy for Large Language Model (LLM)–/Machine Learning (ML)–Extracted Information and Data (VALID) Framework Globally: Quality Assessment of LLM-Derived Prostate Cancer Real-World Data in the UK

ISPOR Europe 2026
POSTER CODE
MSR263
Scan to learn more

Victor Lhoste[1], Arun Sujenthiran[1], Kathi Seidl-Rathkopf[2]; Natalia Viani[1], Marina Kushnir[1]; Nikola Dolezalova[1], Anna Schwarz[2], Amin Hashemian[1], Conor Walsh[1], Alex Enrique[1], Golnessa Mojtahedi[3], Cornelius Thaiss[2], Patrycja Pluta[2], Qianyi Zhang[3]

[1]Flatiron Health, London, UK • [2]Flatiron Health, Berlin, Germany • [3]Flatiron Health, New York, NY, USA

BACKGROUND

Large language models (LLMs) offer a scalable approach to abstracting clinically meaningful data from electronic health records (EHR), but their suitability for generating real-world evidence (RWE) requires rigorous quality assessment. The VALID[1] framework evaluates LLM-derived data quality across three dimensions: variable-level metrics, verification checks, and replication analyses

Objective:
This study applied the VALID framework to a novel, LLM-derived UK prostate cancer (PC) dataset to determine its suitability for generating RWE

METHODS

- **Data source**: All cases of PC spanning initial diagnosis through metastatic castration-resistant disease from the UK-based, EHR-derived, deidentified Flatiron Health Research Database.[2] LLM-derived data (n = 13 ,905) were benchmarked against human-abstracted data (n = 4335) and against national registry data

- **Variables:** LLMs extracted clinically meaningful characteristics, including initial/metastatic diagnosis, castration-resistant status, and treatment. As a reference standard, the same variables were curated via expert human abstraction

- **Statistical methods:** The three dimensions of the VALID[1] framework were used as tools to assess LLM-derived data quality:

Variable-Level Performance Metrics

587 patients underwent double abstraction; one abstractor was randomly designated as reference. F1 scores were computed for the LLM and the second abstractor against this reference; ΔF1 (LLM – abstractor F1) assessed human-level performance

Automated Verification Checks

Prevalence of conflicting or clinically implausible data was assessed at patient and cohort levels

Replication and Benchmarking Analyses

(1) rwOS from metastatic diagnosis was compared between LLM-derived and human-abstracted data in patients receiving ≥1 ARPI- or taxane-containing line in mCRPC;
(2) dataset representativeness was benchmarked against NHS England NDRS estimates and National Prostate Cancer Audit (NPCA) reports

CONCLUSIONS

The **successful application of the VALID[1] framework** to a UK PC cohort demonstrates that LLMs can accurately extract robust, fit-for-purpose data, **enabling faster and highly scalable RWE generation in the UK.**

RESULTS

1. Variable level metrics

Table 1. Variable level metrics

Variable name	Absolute delta F1 score (LLM – abstraction)
Initial diagnosis date*	2.7%
Metastatic diagnosis date*	0.8%
Clinical stage	1.0%
ADT* start date	0.3%

*Note: 30 days tolerance is applied for dates
ADT: Androgen Deprivation Therapy

2. Verification checks

Verification checks assessed the prevalence of conflicting or clinically implausible data points, such as:

- **Implausible treatments:** Confirming radical prostatectomy was appropriately rare in de novo metastatic (0.3%) vs. recurrent cases (15.2%)
- **Logical date sequencing:** Confirming no patients had radiopharmaceutical treatment or CRPC progression recorded prior to initial diagnosis
- **Potential missing data:** Confirming no patients had pre-diagnostic somatic *BRCA* testing, ensuring valid historical germline tests were not inadvertently missed
- **Cohort-level checks:** Verifying that longitudinal outcomes, such as time from HSPC to CRPC and rwOS, are consistent with clinical expectations

3. Replication analyses

3.1 Comparison with expert-human abstraction:

- Median survival time (months, 95% CI) for patients with at least one line of therapy with an ARPI in the metastatic castration-resistant setting (mCRPC) is similar in the LLM-extracted (44.9, 41.5-47.8) and expert human-abstracted (42.8, 38.9-48.0) datasets **(Figure 1)**
- Median survival time (months, 95% CI) for patients with at least one line of therapy with a taxane in the mCRPC setting is similar in the LLM-extracted (37.1, 33.4-42.0) and expert human-abstracted (37.0, 32.9-44.6) datasets **(Figure 2)**

Figure 1. rwOS in patients with at least one line with an ARPI in the mCRPC setting

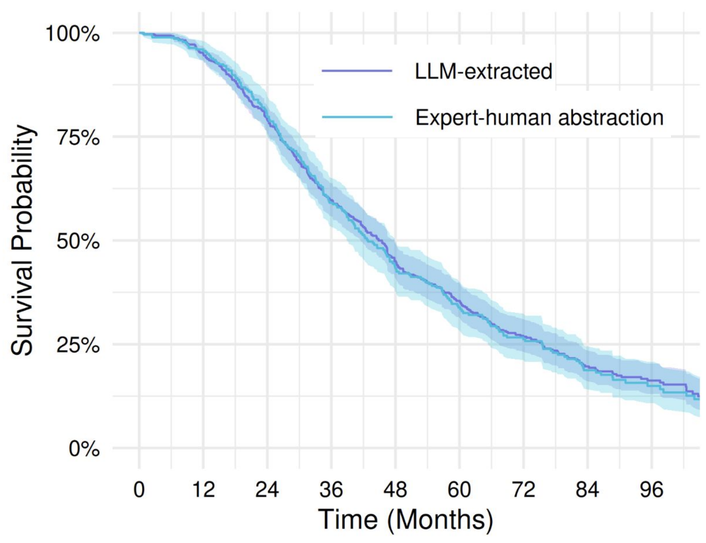
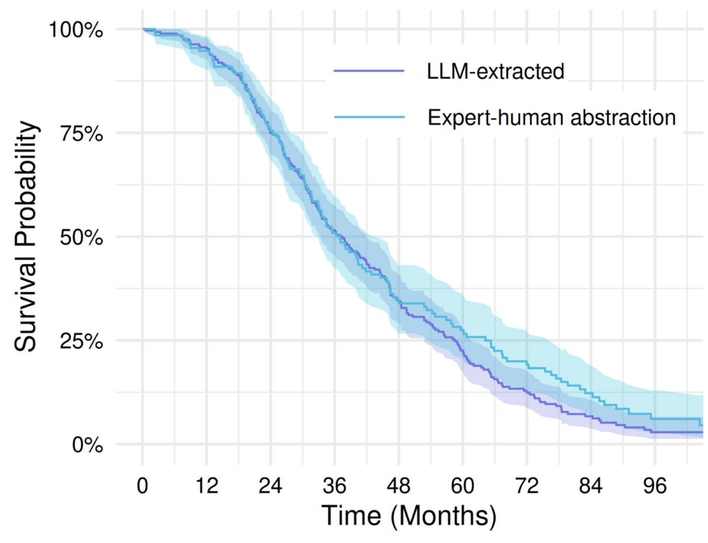


Figure 2. rwOS in patients with at least one line with taxane in the mCRPC setting



3.2 Comparison with national cancer registry

rwOS by stage at diagnosis in the LLM-extracted UK prostate cancer dataset was consistent with NHS England registry[3] estimates (**Table 2**). Longitudinally, the proportion of patients presenting with metastatic disease broadly aligned with published NPCA[4] benchmarks (**Table 3**). Annual radical prostatectomy rates were similar to NPCA[4] estimates, with a slightly higher rate observed in the most recent year (**Table 4**)

Table 2. Overall survival by stage at diagnosis

Stage at diagnosis	NDRS[3] 1-year rwOS	LLM-extracted 1-year rwOS
Stage I	98.4 (98.3-98.5)	99.6 (99.2-100)
Stage II	98.7 (98.6-98.9)	99.5 (98.9-100)
Stage III	97.7 (97.5-97.8)	98.9 (98.5-99.3)
Stage IV	84.3 (84-84.7)	85.4 (84-86.9)

Table 3. Proportion metastatic at diagnosis over time

Time Period	NPCA[4] Benchmark (%)	LLM-Extracted (%)
April 2015-Mar 2016	16.0	18.1
Apr 2016 - Mar 2017	15.9	15.7
Apr 2017 - Mar 2018	16.7	16.8
Apr 2018 - Mar 2019	13.0	15.4
Apr 2019 - Mar 2020	13.0	16.4
Apr 2020 - Mar 2021	17.0	19.5
Apr 2020 - Mar 2021	19.4	19.5
Jan-Dec 2021	16.5	19.0
Jan-Dec 2022	12.5	14.2

Table 4. Radical prostatectomy as percent of new diagnoses

Diagnosis Year	NPCA[4] Benchmark (%)	LLM-Extracted (%)
2019	16.5	15.4
2020	16.9	14.8
2021	14.6	15.1
2022	14.8	15.6
2023	15.9	15.4
2024	16.5	19.9

Future work
Having demonstrated fit-for-purpose performance for UK PC, future work will focus on applying the VALID framework to other cancers and in other countries

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 September 29, 2026 https://flatiron.com/database-characterization
3. NHS England. Cancer Survival in England, cancers diagnosed 2018 to 2022, followed up to 2023. NHS England Digital. Published Apr 9, 2026 https://digital.nhs.uk/data-and-information/publications/statistical/cancer-survival-in-england/cancers-diagnosed-2018-to-2022-followed-up-to-2023
4. National Prostate Cancer Audit (NPCA). State of the Nation Report 2025 available at: https://www.natcan.org.uk/reports/npca-state-of-the-nation-report-2025/

Disclosures
Study sponsored by Flatiron Health, Inc.—an independent member of the Roche Group. Flatiron-affiliated authors were employees of Flatiron Health during the study period

Presenting Author Contact: Victor Lhoste, PhD | victor.lhoste@flatiron.com

Data first presented at ISPOR Europe 2026, Vienna, Austria

Poster MSR263 Monday, 9 November 2026 • 10:30–13:30 • Discussion 12:30–13:30