2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

A Retrospective Study Protocol Benchmarking Natural Language Processing and Computer Vision Models for Named Entity Recognition In Prostate Cancer: Extracting Staging, Histology, and Biomarker Data from Clinical Text (HL7 FHIR mCODE)

Sandeepan Ganguly, MS[1]; Brian Shen, MS[1]; Sunit Jariwala, MD, FAAAAI, FAMIA[2]
[1]Department of Radiation Oncology, Montefiore Einstein Comprehensive Cancer Center, New York, NY, [2]Department of Medicine, Montefiore Health System, New York, NY

ABSTRACT

Purpose: Approximately 80% of clinically actionable oncology data remains siloed in unstructured formats, including clinical narratives and pathology reports. This study validates an end-to-end Natural Language Processing (NLP) pipeline for extracting oncological Named Entities and transforming them into HL7 FHIR resources compliant with the Minimal Common Oncology Data Elements (mCODE) Implementation Guide, enabling computable, interoperable cancer data at scale.
Methods: This retrospective, observational validation study analyzes de-identified longitudinal clinical notes from 50 prostate cancer patients receiving radiation therapy. Two domain experts will independently annotate clinical history & workup summaries using standardized guidelines within a dedicated platform, with inter-annotator agreement assessed using Cohen's-kappa. A commercial NLP and Computer Vision platform will be benchmarked against transformer-based open-source models (BioBERT, SciSpacy, GPT-based architectures) for Named Entity Recognition across mCODE-defined elements including TNM staging, histologic grade, biomarker status, treatment response and 409 other. Extracted entities will be programmatically mapped to Value Set Authority Center terminologies (SNOMED CT, RxNorm, LOINC) via FHIR CodeX specifications. Entity-level performance will be quantified using accuracy with Wilson score 95% confidence intervals. McNemar's test will evaluate discordant pairs between predictions and gold standard to detect significant performance differences. A one-sided binomial test will assess whether models achieve the 95% accuracy threshold. Subgroup analyses will examine performance variation across mCODE data elements.
Results: Primary outcomes include element-wise accuracy with confidence intervals for each mCODE category and comparative performance rankings across models. Secondary outcomes will characterize error patterns, identify concepts with suboptimal extraction, and quantify FHIR mapping fidelity.
Conclusion: This study establishes a rigorous validation framework for clinical NLP pipelines in oncology. Findings will inform technology selection, identify therapeutic targets, and de-risk institutional AI investment. Successful validation will accelerate automated clinical trial screening, real-time quality measure reporting, and population-level cancer surveillance bridging unstructured documentation and actionable, standards-compliant oncology data.

Scan Me

INTRODUCTION

An estimated 80% of clinically actionable oncology information, including staging, histologic grade, biomarker status, and treatment response, exists only as free text, inaccessible to the automated pipelines that drive trial matching, quality measurement, and cancer surveillance. Manual abstraction does not scale: a 50-patient cohort requires 150 to 200 expert hours, prohibitive at the volume of a typical academic center.
The Minimal Common Oncology Data Elements (mCODE) Implementation Guide, developed through the HL7 FHIR CodeX Accelerator, defines a standardized data model for diagnosis, staging, biomarkers, treatment, and outcome, enabling interoperable exchange across EHRs and registries. Realizing this at scale requires a reliable way to convert existing narrative documentation into mCODE-conformant FHIR resources rather than re-abstracting records by hand.
NLP offers a plausible route. Biomedical transformer models (BioBERT, SciSpacy) and general-purpose large language models show promise on clinical entity extraction, and commercial NLP/CV platforms offer a vendor-supported alternative. Missing from the literature is a rigorous, head-to-head comparison of these approaches on mCODE element extraction, carried through to actual FHIR resource generation rather than entity recognition alone.
This study establishes such a framework in a prostate cancer radiation therapy cohort, aiming to: (1) build an expert-adjudicated gold standard; (2) benchmark a commercial platform against BioBERT, SciSpacy, and an open-weight GPT model; (3) quantify entity-level accuracy with rigorous statistics; and (4) assess fidelity of mapping to VSAC terminologies and FHIR mCODE resources.

METHODS AND MATERIALS

Design and Data
This retrospective, observational validation study uses de-identified, longitudinal clinical notes from 50 prostate cancer patients treated with radiation therapy at a single academic center. Clinical history and workup summaries were chosen as the annotation unit since they consistently contain the target diagnosis, staging, biomarker, treatment, and outcome information.
Gold Standard Development
Two domain experts independently labeled each note against standardized guidelines. Inter-annotator agreement was assessed with Cohen's kappa; concepts below a pre-specified agreement threshold prompted guideline refinement and re-annotation until acceptable agreement was reached. Discordant labels were adjudicated by consensus to produce the final gold standard.
Target mCODE Data Elements
Extraction targets span five mCODE categories (Table 1): diagnosis (tumor site, histology), staging (TNM, Gleason score), biomarker (PSA), treatment (RT type, dose), and outcome (disease status).

RESULTS

As this study is presented at the protocol stage, figures below are projected benchmarks, informed by pilot annotation and comparable published NLP tasks, not final outcomes; they will be superseded once full-cohort execution is complete.
Projected Model Performance
The commercial NLP/CV platform is projected to lead at 94% overall accuracy, followed by GPT-oss (91%), BioBERT (88%), and SciSpacy (85%). None is projected to reach the 95% threshold, implying human review would remain necessary before unsupervised downstream use.

Projected Accuracy by mCODE Element
Accuracy is projected to vary by element: TNM stage (92%), histology (89%), treatment (88%), and disease status (85%) perform best, while biomarker (PSA) extraction is weakest at 78%, likely reflecting greater lexical and numerical variability in how lab values are documented.

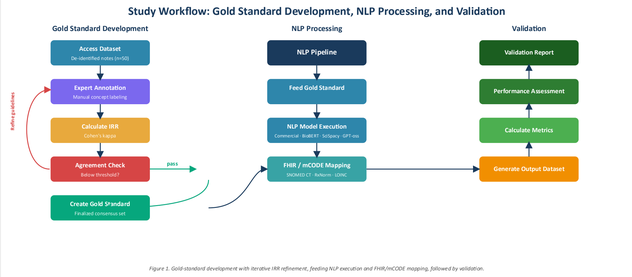

Figure 1. Gold-standard development with iterative IRR refinement, feeding NLP execution and FHIR/mCODE mapping, followed by validation.

Figure 1. Workflow

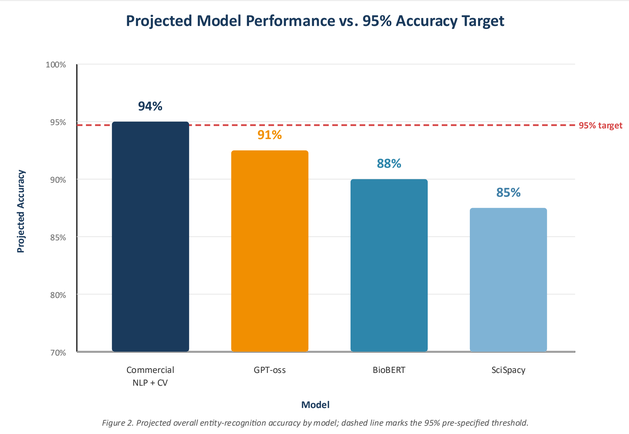

Figure 2. Projected overall entity-recognition accuracy by model; dashed line marks the 95% pre-specified threshold.

Figure 2. Model Performance

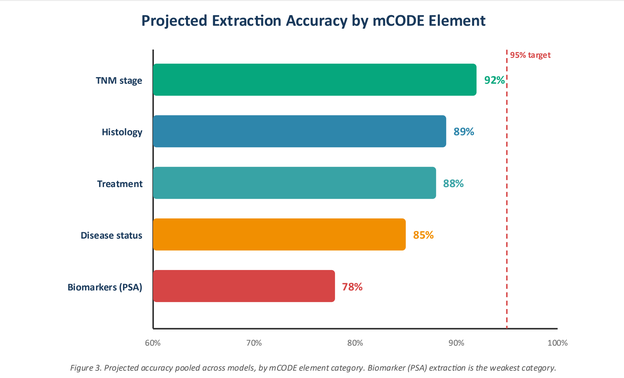

Figure 3. Projected accuracy pooled across models, by mCODE element category. Biomarker (PSA) extraction is the weakest category.

Figure 3. Accuracy Analysis

DISCUSSION

If projections hold, a commercially licensed platform would modestly outperform open-source models, but the margin over GPT-oss is narrow, an important consideration for technology selection given licensing costs versus a locally tunable open-source pipeline.
Anticipated weakness in biomarker extraction is consistent with prior literature: lab values in free text are prone to unit inconsistency, abbreviation variance, and temporal ambiguity. Targeted rule-based normalization layered on general NER may help more than architecture changes alone.
Limitations include a single cancer type and institution (50 patients), possible commercial-platform tuning to institutional templates, an upper bound on measurable accuracy set by inter-annotator agreement, and a 95% threshold that is a design choice rather than an empirically derived requirement. Future work should extend validation to other disease sites and institutions, explore active learning to reduce annotation burden, and monitor model drift if deployed in live clinical workflows.

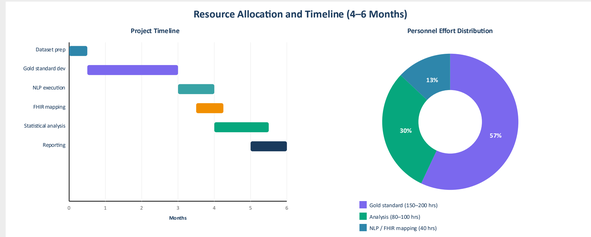

Figure 5. Sequential project timeline (left) and personnel effort distribution (right); gold-standard annotation dominates total effort.

CONCLUSIONS

This protocol establishes a rigorous framework for validating an end-to-end pipeline converting unstructured oncology documentation into mCODE-conformant FHIR data, benchmarking a commercial platform against three open-source models. Projections suggest all approaches fall short of a 95% threshold, with biomarker extraction the consistent weak point. Completing the full 50-patient cohort will test these projections and inform institutional technology selection and investment in clinical NLP infrastructure.

REFERENCES

1. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
2. McNemar Q. Note on the sampling error of the difference between correlated proportions or percentages. Psychometrika. 1947;12(2):153–157.
3. Wilson EB. Probable inference, the law of succession, and statistical inference. J Am Stat Assoc. 1927;22(158):209–212.
4. Lee J, Yoon W, Kim S, Kim D, Kim S, So CH, Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. 2020;36(4):1234–1240.
5. Neumann M, King D, Beltagy I, Ammar W. ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop, Association for Computational Linguistics. 2019:319–327.
6. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Proceedings of NAACL-HLT. 2019:4171–4186.
7. HL7 International. HL7 FHIR Release 4 (R4). Health Level Seven International.
8. HL7 FHIR CodeX Accelerator. FHIR CodeX Implementation Guide.
9. Osterman TJ, Terry M, Miller RS. Improving cancer data interoperability: the promise of the Minimal Common Oncology Data Elements (mCODE) Initiative. JCO Clin Cancer Inform. 2020;4:993–1001.
10. Kehl KL, Elmarakeby H, Nishino M, et al. Assessment of Deep Natural Language Processing in Ascertaining Oncologic Outcomes From Radiology Reports. JAMA Oncol. 2019;5(10):1421–1429.
11. American Society of Clinical Oncology (ASCO) and MITRE Corporation. Minimal Common Oncology Data Elements (mCODE) Implementation Guide.

© Poster Template by Genigraphics® 1.800.790.4001 www.genigraphics.com