

Utility of Oncology Data Standards to Identify Common Data Elements in Breast Cancer Trials' Eligibility Criteria

Adit Anand, MA¹, Karthik Natarajan, PhD¹

¹Department of Biomedical Informatics, Columbia University Irving Medical Center, New York, NY, USA

Background

Delays in patient enrollment are a considerable roadblock in the completion of clinical trials, as estimates indicate that 80% of clinical trials experience delays due to insufficient enrollment and 20% of oncology trials fail as a result(1,2). Given the bottlenecks of trial recruitment, clinical research informatics efforts have focused on automating the identification of eligible patients by feeding in trials' eligibility criteria, the free-text requirements for patients to enroll(3–5). However, this task comes with its set of challenges, as there is no standard computational representation for eligibility criteria due to the lexical variance of criteria across trial protocols(6). Identifying common data elements (CDEs), generalizable variables used to standardize data collection, across trials' eligibility criteria addresses the above gap. Within clinical research informatics, CDEs facilitate the development of interoperable healthcare systems and enable effective secondary usage of healthcare data(7). Given the algorithmic nuances of using eligibility criteria for trial recruitment solutions, uncovering the CDEs in trial criteria would have a similar positive effect(8).

Extracting and mapping the concepts present in eligibility criteria to medical vocabularies is a considerable aspect in curating a set of trial criteria CDEs(9). However, standardized Observational Medical Outcomes Partnership (OMOP) vocabularies struggle to generalize well to cancer concepts. This limitation has prompted efforts to extend the OMOP Common Data Model to oncology data, including the integration of cancer data standards(10). To our knowledge, no current literature uses oncology data standards to identify eligibility criteria CDEs. In this work, we investigate the utility of using oncology vocabularies to generate eligibility criteria CDEs in comparison to using OMOP standardized vocabularies.

Methods

Data Sources

We leverage Clinical Trials Knowledge Base (CTKB), a relational database of discrete criteria and attributes that are derived from processing the eligibility criteria queried from ClinicalTrials.gov, for our analysis(8). CTKB uses Criteria2Query's entity extraction, a rule-based and deep learning information extraction pipeline, to extract clinical entities from the eligibility criteria and assign them to an OMOP domain. Our eligibility criteria dataset covers 1,015 breast cancer trials that are registered on ClinicalTrials.gov as interventional studies, have a study status of "Completed" or "Active, not recruiting", and have the National Cancer Institute listed as a collaborator or sponsor. We focus on the clinical entities that CTKB classifies as one of four OMOP domains: Condition Occurrence, Drug Exposure, Measurement, and Procedure Occurrence.

Concept Mapping

After obtaining our dataset of eligibility criteria entities, we pass these entities through two separate concept mapping pipelines. The first pipeline ("Baseline") maps entities to OMOP concepts found in the standard vocabularies specified. The second pipeline ("Oncology") maps entities to OMOP concepts found in cancer vocabularies. The vocabularies used in each pipeline are specified in Table 1. We employ the Hecate API, a semantic search engine for the Observational Health Data Sciences and Informatics (OHDSI) vocabulary, to map entities in each pipeline(11). We exclude any concept mappings with a semantic score below 0.6 in our analysis.

CDE Analysis

For each OMOP domain, we determine the five most frequent CDEs and their frequency across all trials. We do not consolidate semantically similar CDEs across OMOP domains into a single reported CDE but rather treat them as separate CDEs to adhere with CTKB entity extraction methodology (e.g., Procedure Occurrence concepts we categorize as “Chemotherapy” are reported separately from Drug Exposure concepts we categorize as “Chemotherapy”). Details about how OMOP concepts are rolled up into reported CDEs can be found in Table 1 for the Baseline and Oncology pipelines respectively.

Table 1. Details on each concept mapping pipeline applied to the entities extracted from the breast cancer trials’ eligibility criteria. Each row corresponds to a specific OMOP domain and specifies the corresponding vocabulary used in each pipeline. The concept class used to classify concepts into CDEs is specified in each row.

	Baseline		Oncology	
	Vocabularies	Classification Strategy	Vocabulary	Classification Strategy
Condition	SNOMED	MedDRA	HemOnc	N/A
Drug	RxNorm	ATC 3rd Level	HemOnc	HemOnc Modality
Measurement	LOINC	LOINC Component	NAACCR	NAACCR Measurement Variable
Procedure	SNOMED	CPT4 Hierarchy	NAACCR	NAACCR Procedure Variable

Results

Table 2 shows the five most frequent CDEs for each OMOP domain identified through each concept mapping pipeline. For the Drug domain, we observe that several of the Baseline CDEs can be categorized as an Oncology CDEs (e.g., “Other Neoplastic Agents” would fall under “Chemotherapy”), which reflects HemOnc’s hierarchy being designed for cancer treatment regimens. Furthermore, the Oncology CDEs for the Condition domain are more cancer-specific compared to their respective Baseline CDEs. Additionally, the sets of Measurement CDEs overlap in their contents and comparable levels of granularity. We also observe overlap across the Procedure domain, although drug-based procedures could not be mapped to the NAACCR vocabulary.

The results from the Oncology pipeline showcase the strengths and shortcomings of certain oncology vocabularies. CDEs from the Condition, Drug, and Measurement domains demonstrate that using these data standards for concept mapping captures more granular oncology context. However, high-frequency Procedure CDEs from the Baseline pipeline (e.g., “Administration of antineoplastic agent”, “Adjuvant drug therapy”) are not reflected in the Oncology pipeline’s Procedure CDEs. This indicates that oncology vocabularies are not suitable by themselves to be applied to lexically diverse text such as trial eligibility criteria.

Table 2. Each set of CDEs identified by mapping extracted entities to the concepts in either standard OMOP vocabularies or oncology vocabularies. Each row corresponds to CDE under a specific OMOP domain and includes the CDE concept and count of breast cancer trials the concept is found in.

	Baseline		Oncology	
	CDE Name	Number of Trials	CDE Name	Number of Trials
Condition	Breast carcinoma	689 (67.882%)	Breast cancer	811 (79.901%)
	Pregnancy	637 (62.759%)	Skin cancer	417 (41.084%)
	Carcinomatosis	511 (50.345%)	Hepatocellular carcinoma	319 (31.429%)
	Malignant neoplasm NOS	394 (38.818%)	Renal cell carcinoma	304 (29.951%)
	Hepatic Failure	332 (32.709%)	CNS cancer	294 (28.966%)
Drug	Cytotoxic Antibiotics and Related Substances	272 (26.798%)	Endocrine Therapy	554 (54.581%)
	Hormone Antagonists and Related Agents	242 (23.842%)	Chemotherapy	337 (33.202%)
	Contraceptives for Topical Use	195 (19.212%)	Targeted Therapy	210 (20.690%)
	Monoclonal Antibodies and Antibody Drug Conjugates	175 (17.241%)	Supportive Therapy	201 (19.803%)
	Other Antineoplastic Agents	168 (16.552%)	Immunosuppressive Therapy	125 (12.315%)
Measurement	Bilirubin	508 (50.049%)	Bilirubin	510 (50.246%)
	Platelets	505 (49.754%)	Creatinine	298 (29.360%)
	Neutrophils/ Leukocytes	310 (30.542%)	HER2 Biomarker	159 (15.665%)
	Performance Status	287 (28.276%)	Karnofsky Performance Status	129 (12.709%)
	Hemoglobin	235 (23.153%)	HIV Status	87 (8.571%)
Procedure	Administration of antineoplastic agent	679 (66.900%)	Mastectomy	143 (14.089%)
	Adjuvant drug therapy	249 (24.532%)	Oophorectomy	132 (13.005%)
	Transcatheter Diagnostic Radiology (Diagnostic Imaging) Procedures	192 (18.916%)	Hysterectomy	98 (9.655%)
	Mastectomy Procedures	107 (10.542%)	Lumpectomy	46 (4.532%)
	Hysterectomy	97 (9.557%)	Breast Implant	33 (3.251%)

Conclusion

In this preliminary work, we investigated the feasibility of using oncology data standards integrated into OHDSI's vocabulary to identify CDEs in breast cancer trials' eligibility criteria. Our findings highlight that we can identify more granular cancer CDEs by leveraging oncology vocabularies, but these vocabularies do not capture the wider breadth of CDEs pertinent to oncology trial recruitment. For future directions, we aim to explore how OHDSI's standardized and oncology vocabularies can be jointly used to inform how we curate CDEs. Furthermore, we plan to apply this methodology to trials for other types of cancer to evaluate the generalizability of our findings. Finally, subsequent work could employ more aspects of the OHDSI vocabularies' relationships to identify eligibility criteria CDEs in a more algorithmic manner.

References

1. Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard clinical trial protocols. *J Biomed Inform.* 2013 Oct 1;46(5):805–13. doi:10.1016/j.jbi.2013.06.001
2. Allen J, Bergman K, Bruinooge S, Davis A, Downs CG, Ferris A, et al. Barriers to Patient Enrollment in Therapeutic Clinical Trials [Internet]. American Cancer Society Cancer Action Network; 2018 Apr. Available from: <https://www.fightcancer.org/sites/default/files/National%20Documents/Clinical-Trials-Landscape-Report.pdf>
3. Criteria2Query 3.0: Leveraging generative large language models for clinical trial eligibility query generation - ScienceDirect [Internet]. [cited 2025 Jul 1]. Available from: <https://www.sciencedirect.com/science/article/pii/S1532046424000674>
4. Syed M, Hamidi M, Bikkanuri M, Dierschke NA, Katragadda HV, Zozus M, et al. Translating evidence into practice: adapting TrialGPT for real-world clinical trial eligibility screening. *J Am Med Inform Assoc.* 2026 Apr 1;33(4):909–13. doi:10.1093/jamia/ocag006
5. Lin J, Xu H, Wang Z, Wang S, Sun J. Panacea: A foundation model for clinical trial search, summarization, design, and recruitment [Internet]. arXiv; 2024 [cited 2025 Dec 17]. Available from: <http://arxiv.org/abs/2407.11007> doi:10.48550/arXiv.2407.11007
6. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. ElixR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc.* 2011 Dec 1;18(Supplement_1):i116–24. doi:10.1136/amiajnl-2011-000321
7. Sheehan J, Hirschfeld S, Foster E, Ghitza U, Goetz K, Karpinski J, et al. Improving the value of clinical research through the use of Common Data Elements. *Clin Trials.* 2016 Dec 1;13(6):671–6. doi:10.1177/1740774516653238
8. Liu H, Chi Y, Butler A, Sun Y, Weng C. A knowledge base of clinical trial eligibility criteria. *J Biomed Inform.* 2021 May 1;117:103771. doi:10.1016/j.jbi.2021.103771
9. Yuan C, Ryan PB, Ta C, Guo Y, Li Z, Hardin J, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc.* 2019 Apr 1;26(4):294–305. doi:10.1093/jamia/ocy178
10. Belenkaya R, Gurley MJ, Golozar A, Dymshyts D, Miller RT, Williams AE, et al. Extending the OMOP Common Data Model and Standardized Vocabularies to Support Observational Cancer Research. *JCO Clin Cancer Inform.* 2021 Jan;(5):12–20. doi:10.1200/CCI.20.00079
11. OHDSI/Hecate [Rust] [Internet]. Observational Health Data Sciences and Informatics; 2026 [cited 2026 Jun 5]. Available from: <https://github.com/OHDSI/Hecate>