

Preliminary Evaluation of Common Data Elements Coverage of Oncology Clinical Trials' Eligibility Criteria within OMOP

Adit Anand, MA¹, Karthik Natarajan, PhD¹

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

Background

Clinical trials serve as a gold standard for data that shapes the healthcare landscape in oncology(1,2). However, a considerable bottleneck for clinical trials is failure to recruit the target number of participants(3). There are growing efforts to incorporate real-world health data into trial recruitment tools(4,5), and evaluating the data's fitness for use benefits the development of these solutions(6). One aspect of this evaluation involves identifying common data elements (CDE) present in clinical trials since CDEs assist with standardizing how clinical data is represented(7). In this study, we identified CDEs present in the eligibility criteria of clinical trials and determined their prevalence in the Columbia University Irving Medical Center (CUIMC) OMOP database.

Methods

For this analysis, we identified oncology CDEs across clinical trials for brain cancer, breast cancer, and prostate cancer. Using the ClinicalTrials.gov API, we curated a collection of clinical trials for each cancer type that were posted between the calendar years of 2010 to 2024. For each cancer type, we queried ClinicalTrials.gov with the cancer type verbatim (e.g., breast cancer results were curated using the search term "breast cancer").

We leveraged MedSpaCy(8), a toolkit for clinical natural language processing, to extract clinical entities present in the eligibility criteria of each clinical trial. We then mapped these entities to corresponding OMOP concepts. For each set of cancer trials, we identified the five most common data elements across the Condition Occurrence, Drug Exposure, Measurement, and Procedure Occurrence tables.

Using the CDEs' corresponding OMOP concepts, we computed the coverage of these CDEs across corresponding cohorts that we defined for each cancer type. The cohorts were defined using OHDSI's ATLAS tool. We leverage CUIMC's OMOP database to generate the results. The code used to extract data elements from ClinicalTrials.gov and map the entities to OMOP concepts along with the cohort definitions can be found in the following GitHub repository (<https://github.com/adit-anand/oncology-clinical-trials-cdes>).

Results

Table 1 shows the five most recurring CDEs in eligibility criteria for brain cancer trials across four OMOP domain tables. Table 2 and Table 3 presents similar CDE information for breast cancer and prostate cancer trials, respectively.

We observe the Measurement domain CDEs present in clinical trials are consistently populated in CUIMC OMOP for cohort subjects, with the exception of oncology-specific CDEs (i.e., estrogen and Karnofsky Performance Status). Furthermore, the lowest levels of CDE prevalence in CUIMC OMOP occur within the Drug Exposure domain. The lowest alignment between clinical trials and OMOP records can be seen for CDEs in the Procedure domain, as the OMOP prevalences of several CDEs are significantly lower than their corresponding clinical trial prevalence.

Clinical Domain	Common Data Element	Trial Count (n = 2,124)	Cohort Subject Count (n = 45,664)
Condition	Carcinoma of Breast	154 (7.250%)	703 (1.630%)
	Gilbert's Syndrome	146 (6.874%)	31 (0.068%)
	Opitz-Frias Syndrome	141 (6.638%)	0 (0.000%)
	Psychiatric Disorder	138 (6.497%)	16,617 (36.390%)
	Squamous Cell Carcinoma of Skin	130 (6.121%)	426 (9.329%)
Drug	Temozolomide	194 (9.134%)	1,949 (4.268%)
	Epoetin Alfa	52 (2.448%)	839 (1.837%)
	Interferon Alfa-2b	46 (2.166%)	33 (0.072%)
	Mitomycin	37 (1.742%)	125 (0.273%)
	Pembrolizumab	28 (1.312%)	376 (0.823%)
Measurement	Bilirubin	578 (27.213%)	33,544 (73.458%)
	Karnofsky Performance Status	449 (21.394%)	2,794 (6.119%)
	Neutrophil Count	440 (20.716%)	33,619 (73.623%)
	Hemoglobin	415 (19.539%)	37,801 (82.781%)
	Platelet Count	385 (18.126%)	37,700 (82.560%)
Procedure	Chemotherapy	705 (33.192%)	5,636 (12.342%)
	Implantation	143 (6.733%)	7,441 (16.295%)
	Whole Brain Radiation Therapy	128 (6.026%)	0 (0.000%)
	Oophorectomy	89 (4.190%)	857 (1.877%)
	Hormone Therapy	53 (2.495%)	556 (1.218%)

Table 1. Common data elements found in the eligibility criteria of brain cancer clinical trials across different OMOP clinical domains. Each CDE has a reported number of trials the CDE is found in and number of cohort subjects with an OMOP record of the corresponding CDE.

Clinical Domain	Common Data Element	Trial Count (n = 8,048)	Cohort Subject Cohort (n = 79,405)
Condition	Carcinoma of Breast	5,915 (73.497%)	15,152 (19.082%)
	Autoimmune Disease	798 (9.916%)	5,624 (7.083%)
	Squamous Cell Carcinoma of Skin	680 (8.449%)	977 (1.230%)
	Gilbert's Syndrome	641 (7.965%)	48 (0.060%)

	Opitz-Frias Syndrome	574 (7.132%)	0 (0.000%)
Drug	Gonadorelin	242 (3.007%)	0 (0.000%)
	Pertuzumab	199 (2.473%)	500 (0.630%)
	Mitomycin	166 (2.063%)	178 (0.224%)
	Pembrolizumab	164 (2.038%)	92 (0.116%)
	Epoetin Alfa	156 (1.938%)	970 (1.222%)
		Bilirubin	1,720 (21.372%)
Measurement	Hemoglobin	1,313 (16.315%)	60,299 (75.939%)
	Neutrophil Count	1,212 (15.060%)	53,358 (67.197%)
	Platelet Count	1,037 (12.885%)	60,217 (75.835%)
	Estrogen	894 (11.108%)	6,067 (7.641%)
		Chemotherapy	3,831 (47.602%)
Procedure	Excision of Breast Tissue	734 (9.120%)	25,965 (32.699%)
	Oophorectomy	502 (6.238%)	3,204 (4.035%)
	Hormone Therapy	247 (3.069%)	859 (1.082%)
	Core Needle Biopsy	206 (2.560%)	4,608 (5.803%)

Table 2. Common data elements found in the eligibility criteria of breast cancer clinical trials across different OMOP clinical domains. Each CDE has a reported number of trials the CDE is found in and number of cohort subjects with an OMOP record of the corresponding CDE.

Clinical Domain	Common Data Element	Trial Count (n = 3,687)	Cohort Subject Count (n = 52,289)
Condition	Carcinoma of Prostate	562 (15.243%)	3,102 (5.932%)
	Gilbert's Syndrome	325 (8.815%)	62 (0.119%)
	Spinal Cord Compression	260 (7.052%)	1,230 (2.352%)
	Squamous Cell Carcinoma of Skin	257 (6.970%)	1,283 (2.454%)
	Autoimmune Disease	207 (5.614%)	1,701 (3.253%)
Drug	Gonadorelin	652 (17.684%)	0 (0.000%)
	Liothyronine	109 (2.956%)	113 (0.216%)
	Sipuleucel-T	98 (2.658%)	37 (0.071%)
	Prednisone	65 (1.763%)	7,255 (13.875%)

	Polysorbate 80	56 (1.519%)	16 (0.031%)
Measurement	Bilirubin	848 (23.000%)	37,814 (72.317%)
	Hemoglobin	697 (18.904%)	39,376 (75.305%)
	Neutrophil Count	585 (15.867%)	33,397 (63.870%)
	Platelet Count	502 (13.615%)	39,275 (75.111%)
	Creatinine	427 (11.581%)	44,343 (84.804%)
	Procedure	Chemotherapy	1,228 (33.306%)
Brachytherapy		233 (6.320%)	704 (1.346%)
Transurethral Prostatectomy		183 (4.963%)	2,381 (4.554%)
Implantation		118 (3.200%)	395 (0.755%)
Dissection of Lymph Node		75 (2.034%)	0 (0.000%)

Table 3. Common data elements found in the eligibility criteria of prostate cancer clinical trials across different OMOP clinical domains. Each CDE has a reported number of trials the CDE is found in and number of cohort subjects with an OMOP record of the corresponding CDE.

Conclusion

In this preliminary work, we identify CDEs present in the eligibility criteria of three cancer types and evaluate how well these CDEs are captured within the CUIMC OMOP. We find that there is variance across domain tables in how well eligibility criteria CDEs are captured. Our findings across clinical domains can inform downstream usage of these CDEs such as efforts by the Clinical Trial Working Group to determine best practices for representing clinical trials in OMOP(9). Our immediate next steps will be to evaluate the quality of entity extraction from clinical trials. In future work, we intend to perform this CDE analysis across multiple OMOP institutions to identify site-specific and general trends. Furthermore, we will aim to leverage OMOP's comprehensive hierarchies for the Condition and Drug domains to perform more comprehensive CDE analysis (i.e., grouping granular OMOP concepts into broader CDE categories). Finally, future work could use more robust computational approaches for extracting and mapping medical entities to OMOP to generate more representative eligibility criteria CDEs.

References

1. Masic I, Miokovic M, Muhamedagic B. Evidence Based Medicine - New Approaches and Challenges. *Acta Inform Medica*. 2008;16(4):219.
2. Hansson SO. Why and for what are clinical trials the gold standard? *Scand J Public Health*. 2014 Mar 1;42(13_suppl):41–8.
3. Peterson JS, Plana D, Bitterman DS, Johnson SB, Aerts HJWL, Kann BH. Growth in eligibility criteria content and failure to accrue among National Cancer Institute (NCI)-affiliated clinical trials. *Cancer Med*. 2023;12(4):4715–24.
4. Jin Q, Wang Z, Floudas CS, Chen F, Gong C, Bracken-Clarke D, et al. Matching patients to clinical trials with large language models. *Nat Commun*. 2024 Nov 18;15(1):9074.
5. 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>

6. Raman SR, O'Brien EC, Hammill BG, Nelson AJ, Fish LJ, Curtis LH, et al. Evaluating fitness-for-use of electronic health records in pragmatic clinical trials: reported practices and recommendations. *J Am Med Inform Assoc*. 2022 May 1;29(5):798–804.
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.
8. Eyre H, Chapman AB, Peterson KS, Shi J, Alba PR, Jones MM, et al. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python [Internet]. arXiv; 2021 [cited 2025 Jul 1]. Available from: <http://arxiv.org/abs/2106.07799>
9. projects:workgroups:clinicalstudy [Observational Health Data Sciences and Informatics] [Internet]. [cited 2025 Jul 1]. Available from: <https://www.ohdsi.org/web/wiki/doku.php?id=projects:workgroups:clinicalstudy>