Title:

Establishing the Automatic Identification of Clinical Trial Cohorts from Electronic Health Records by Matching Normalized Eligibility Criteria and Patient Clinical Characteristics

Y. Mai[1*], K. Lee[1*], Z. Liu*[1], K. Raja [1*], M. K. Higashi[1], T. Jun[1], M. Ma[1], T. Wang[1], L. Ai[1], E. Calay[1], W. Oh[1,2], E. Schadt[1,2], X. Wang[1]

[1]Sema4, Stamford, 333 Ludlow Street, CT 06902, USA.

[2]Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl, New York, NY 10029, USA.

[*]Equal contributions.

Abstract

Objective

The use of electronic health records (EHRs) holds promising potential to enhance clinical trial activities. However, the identification of eligible patients within EHRs presents considerable challenges. Our objective was to develop an eligibility criteria phenotyping pipeline that would identify patients with matching clinical characteristics from EHRs.

Material and methods

In this study, we utilized clinical trial eligibility criteria from clinicaltrial.gov and patients' EHR datasets from the Sema4 data warehouse, which include multiple heath provider datasets. To ensure computability and queryability, the eligibility criteria attributes and clinical characteristics in EHRs were normalized using four national standard terminologies, LIONC, ICD-9-CM, ICD-10-CM, and CPT, along with four in-house knowledge bases containing procedures, medications, biomarkers, and diagnosis modifiers. The process involved a semi-automated approach incorporating rule-based, pattern recognition, and manual annotation methods. The quality of machine-normalized criteria attributes was accessed using Cohen's Kappa coefficient on randomly selected criteria, and the accuracy of our matching between normalized criteria and patient clinical characteristics was evaluated using precision, recall, and F1 score on randomly selected patients.

Results

A total of 640 unique eligibility criteria attributes were identified, covering various medical conditions, including five types of cancer (non-small cell lung cancer, small cell lung cancer, prostate cancer, breast cancer, and multiple myeloma), two autoimmune diseases (ulcerative colitis and Crohn's disease), one metabolic disorder (non-alcoholic steatohepatitis), and a rare disease (sickle cell anemia). Among these attributes, 367 eligibility criteria attributes were normalized. 174 were encoded with standard

terminologies and 193 were normalized using the in-house reference tables. The agreement between automated and manually annotated normalized codes was found to be 0.82 and matching between eligibility criteria attribute and patient clinical information achieved a high F1-score of 0.94.

Conclusion

We established a clinical phenotyping pipeline facilitating effective communication between the eligibility criteria and EHR. The pipeline demonstrated its generalizability by being applied to EHR data from different institutes. Our pipeline shows the potential to significantly enhance the utilization of EHRs in clinical trial activities and improve patient matching and selection processes, thereby advancing clinical research and patient outcomes.

Keywords: Eligibility criteria phenotyping, Electronic Healthcare Records, cohort identification, clinical trials, eligibility criteria attribute normalization

Introduction

Patient recruitment and retention are highly challenging in clinical trials [1]. Inadequate patient accrual is a major reason for the failure of clinical trials. It can be caused by the disease or trial-specific difficulties, such as a small sample size for rare disease clinical trials [2,3]. Additionally, the competitive market, lack of knowledge, uncertainties of patients about being a study subject[3,4], and rigid protocols that may exclude a large portion of the target population[5,6] can impede recruitment efforts. Identifying a potential population meeting the essential eligibility criteria can accelerate subject matching and facilitate the review of the feasibility of a clinical trial protocol. However, the traditional approach to interpreting and evaluating medical records on a case-by-case basis is tedious and almost impossible for a large cohort selection. An alternative solution is to leverage computer-assisted approaches and mine data from EHRs for quick prescreening of the patient's eligibility in clinical trials. This approach has proven to be an effective method, offering efficiency and accuracy in eligible cohort identification [7-11].

Electronic clinical phenotyping involves extracting relevant clinical phenotypes and patients' characteristics from large datasets [12]. Clinical phenotyping has gained significant attention in both precision medicine and population-based medicine for applications such as cohort selection for clinical predictive modeling, clinical trial cohort identification, and healthcare quality measurement [13]. The transformation of clinical trial eligibility criteria into computer-interpretable representations has facilitated the identification of clinical phenotypes necessary for various applications, including cohort selection[2,4,5,9,11]. -However, automated clinical phenotyping from EHR data presents considerable challenges [14]. Earlier studies have focused on parsing clinical trial eligibility criteria into computer-interpretable representations [15-17] to facilitate trial protocol design, automated cohort selection, and collaborative clinical research [18-25]. Expression and query languages such as Arden Syntax[23,26], Guideline Expression Language Object-Oriented (GELLO)[27], ECLECTIC[28], and Clinical Trail Markup Language[29,30], use a syntax similar to the computer programming languages for representing eligibility

criteria in a computer interpretable format. Template-based approaches such as Eligibility Rule Grammar and Ontology (ERGO)[31] and Eligibility Criteria Extraction and Representation (EliXR)[32], transform eligibility criteria into computable representations. The computable representations can be applied in SPARQL queries, Web Ontology Language (OWL) DL queries, and SQL queries to automate clinical phenotyping. Criteria2SQL and similar works represent EC as SQL queries[15-17]. Certain approaches structure eligibility criteria with Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM)[33,34]. However, existing approaches that link eligibility criteria to EHR face two limitations. Firstly, the syntax of the expression and query languages, as well as intermediate representation, relies on natural language representation, which may result in inaccurate or incomplete data retrieval due to abbreviations and potential typos in clinical terms. Secondly, the accessibility of syntax is not straightforward. These limitations can be overcome by building an advanced intermediate representation of normalized and standardized clinical concepts that can be easily implemented through SQL queries.

In this study, we implemented eligibility criteria phenotyping pipeline, comprising three components. Firstly, we developed a rule-based knowledge engineering component to annotate the extracted eligibility criteria attributes (previously generated using a natural language processing (NLP)-assisted approach) into a computable and customizable granularity from EHRs. Secondly, we normalized the heterogeneity of clinical expressions in the annotated eligibility criteria-attributes and EHRs to predefined medical concepts from standard terminologies and four in-house knowledge bases (procedures, medications, biomarkers, and diagnosis modifiers). Thirdly, we constructed a knowledge base of computable criteria attributes to match patients to clinical trials. The extracted eligibility criteria-attributes in the knowledge base can be utilized for various purposes, including cohort selection, trial protocol design, and many more.

Materials and methods

Overview

For this study, we utilized the eligibility criteria attributes extracted from a total of 3,475 clinical trials. Among these, 3,281 clinical trials, recruiting patients with non-small cell lung cancer, prostate cancer, breast cancer, multiple myeloma, ulcerative colitis, and Crohn's disease were previously analyzed leveraging a deep-learning based NLP technique (manuscript submitted, https://preprints.jmir.org/preprint/50800). An additional 194 trials recruiting small cell lung cancer, non-alcoholic steatohepatitis, and sickle cell anemia were analyzed prior to clinical phenotyping. All extracted eligibility criteria attributes were categorized into ten clinical domains: condition, procedure, laboratory test, therapy, biomarker, observation, diagnosis modifier, line of therapy, vital sign, and demographic. Each eligibility criteria attribute consisted of a name, a group (35 in total), and a value. We also categorized patients' clinical characteristics retrieved from EHR under the same ten clinical domains used for eligibility criteria attributes. We annotated and/or normalized both the eligibility criteria attributes and clinical characteristics to establish the mapping between eligibility criteria and EHR data. The mapped eligibility criteria attributes and clinical characteristics were then saved in a knowledge base for further analysis and reference. The study is covered under IRB-17-01245 approved by the Program for the Protection of Human Subjects at the Mount Sinai School of Medicine.

Data Sources

We collected data from two primary sources: ClinicalTrials.gov (https://clinicaltrials.gov/) and EHR data obtained from Sema4 data warehouse, which include the Mount Sinai Data Warehouse (MSDW) and VieCure, a next-generation clinical decision support (CDS) platform for cancer care (https://www.viecure.com/). The EHR data consisted of various types of information including patient demographics, vital signs, medical histories, diagnoses, medications, laboratory test results, immunization dates, allergies, and radiology images. These data were sourced from the billing documents, progress

notes, radiology reports, pathology reports, operative reports, and discharge summaries collected from all Mount Sinai-associated hospitals and healthcare centers in the Greater New York City area. The data utilization followed the Health Insurance Portability and Accountability Act (HIPAA) regulations.

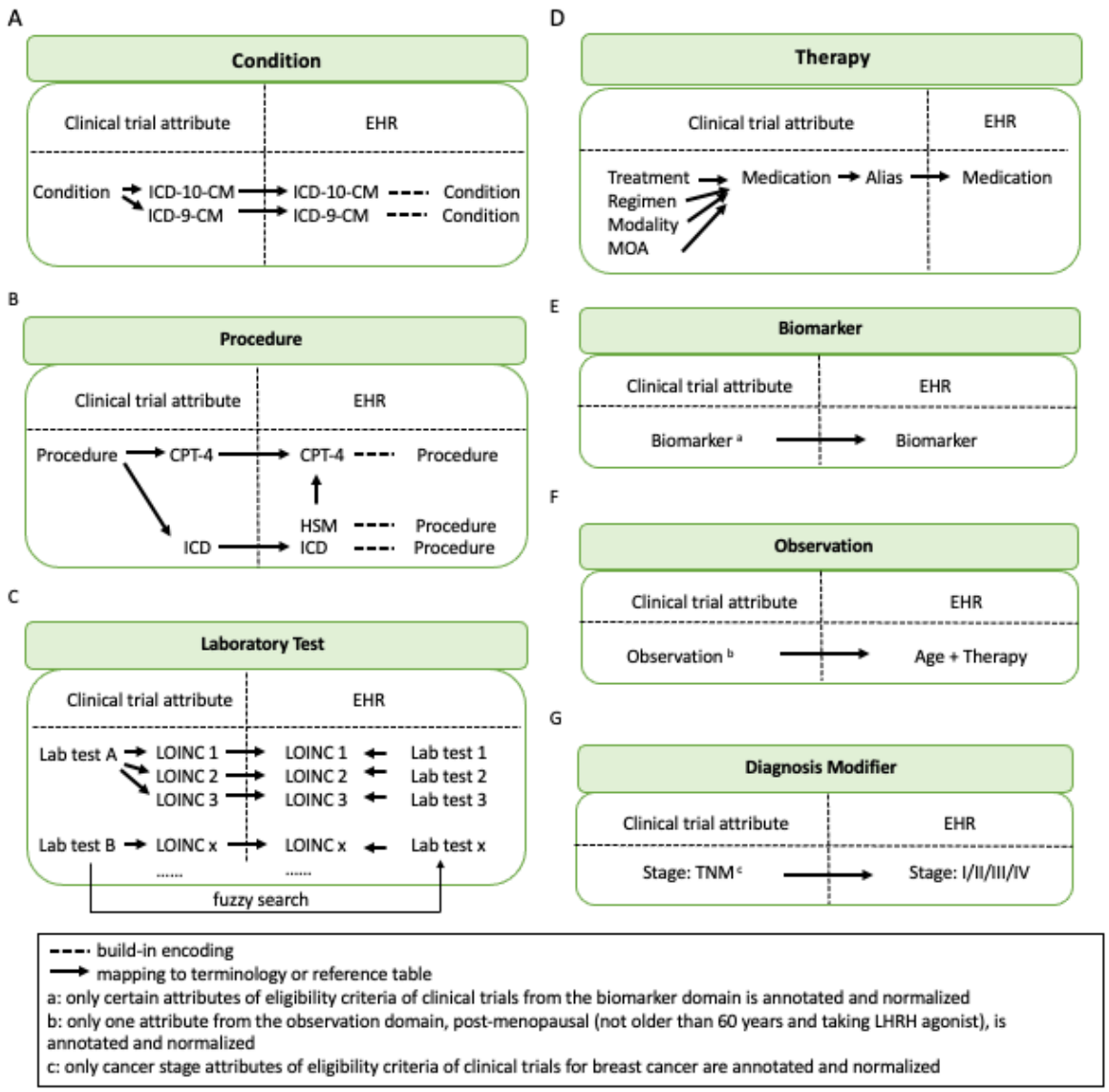
Rule-based knowledge engineering

We employed rule-based knowledge engineering to annotate eligibility criteria attributes and clinical characteristics from EHR within the therapy domain. Standard resources and in-house knowledge bases were utilized for this purpose. The eligibility criteria included five therapy-related terms: (i) treatment (e.g. neoadjuvant therapy), (ii) regimen (e.g. AC and TCH), (iii) modality (e.g. chemotherapy, immunotherapy), (iv) mechanism of action (MOA) (e.g. checkpoint inhibitor, EGFR inhibitor, androgen deprivation therapy), and (v) medication (e.g. platinum-based drugs). Treatment, regimen, modality, and MOA were mapped to specific medications using dedicated resources such as Cancer Alteration Viewer (CAV) and disease treatment guidelines (see Multimedia Appendix 1 for details). For example, the treatment, neoadjuvant chemotherapy for non-small cell lung cancer is mapped to medications such as cisplatin and vinorelbine. The regimen, AC-T for breast cancer is mapped to medications such as doxorubicin hydrochloride, cyclophosphamide, and paclitaxel. MOA, anti-androgen for prostate cancer is mapped to several medications including bicalutamide, flutamide, nilutamide, apalutamide, darolutamide, enzalutamide, and abiraterone.

For laboratory tests, biomarkers, and observations, we annotated attribute names and values. The groups were added to the eligibility criteria attributes before saving them to the knowledge base (Fig. 1). Certain biomarkers within the eligibility criteria attributes (e.g., HER2 R678Q) do not need further annotation while certain biomarkers (e.g., EGFR mutations sensitized to tyrosine kinase inhibitor) need annotation prior to mapping. We annotated such biomarkers with all possible mentions from the literature and examples mentioned in eligibility criteria (e.g., L858R in exon 21, L861Q in exon 21, in-frame

deletions in exon 19) to ensure comprehensive coverage. Medication classes (e.g. LHRH agonist) were annotated with corresponding medication. For example, one of the eligibility criteria attributes, "post-menopausal not older than 60 years and taking LHRH agonist" was annotated to "post-menopausal not older than 60 years and taking goserelin, leuprolide, or other LHRH agonist".

Fig. 1. Clinical trial eligibility criteria phenotyping. (A) Eligibility criteria attributes from the condition domain are annotated and normalized and mapped to clinical characteristics in EHR. (B) Eligibility criteria attributes from the procedure domain and clinical characteristics in EHR are annotated and normalized. Annotated and normalized attributes of eligibility criteria of clinical trials are mapped to normalized clinical characteristics in EHR. (C) Eligibility criteria attributes from the laboratory test domain are annotated and mapped to clinical characteristics in EHR from the laboratory test domain. Clinical characteristics in EHR from the laboratory test domain are annotated and normalized. Attributes of eligibility criteria of clinical trials are normalized through mapping to annotated and normalized clinical characteristics in EHR. (D) Eligibility criteria attributes from the therapy domain are annotated, normalized, and mapped to clinical characteristics in EHR. (E) Certain eligibility criteria attributes from the biomarker domain are annotated, normalized, and mapped to clinical characteristics in EHR. (F) Certain eligibility criteria attributes from the demographic domain are annotated, normalized, and mapped to clinical characteristics in EHR. (G) Certain eligibility criteria attributes from the diagnosis modifier domain are annotated, normalized, and mapped to clinical characteristics in EHR.



Normalization of clinical attributes and clinical characteristics

Eligibility criteria attributes and clinical characteristics within the seven clinical domains, condition, procedure, laboratory test, therapy, biomarker, observation, and diagnosis modifier were normalized using standard resources such as ICD-9-CM (International Classification of Diseases 9th Revision Clinical Modification) and ICD-10-CM (International Classification of Diseases 10th Revision Clinical Modification), and in-house knowledge bases (Fig. 1). While normalization of clinical attributes and clinical characteristics is simple and straightforward within the condition, normalization of clinical attributes and clinical characteristics within the procedure, laboratory tests, and therapy were challenging.

Eligibility criteria attributes within the procedure domain were normalized using Current Procedural Terminology (CPT) 4th Edition, and the standard terminology for procedures from ICD-9-CM or ICD-10-CM (Fig. 1B). The procedures mentioned in EHR are from two sources, the post-surgery documentation system from the EPIC database and Horizon Surgical Manager (HSM). The procedures from EPIC are either encoded with CPT, ICD-9-CM, or ICD-10-CM, or not encoded. We mapped the procedures without encoding in EPIC to CPT using the online CPT bioportal (https://bioportal.bioontology.org/ontologies/CPT). The procedures from HSM are encoded with HSM code. We created an in-house knowledge base to map HSM code to CPT.

Eligibility criteria attributes and clinical characteristics from EHR within the laboratory test domain were normalized using LOINC codes (https://loinc.org/) (Fig. 1C). The system (e.g., serum), quantity (e.g., molar), time (e.g., mol/24h), type of scale (e.g., quantitative), and type of method (e.g., immunoassay) from a laboratory test were used for mapping it to the best LOINC code. For each laboratory test from the eligibility criteria, we performed a fuzzy search to retrieve a list of related laboratory tests from EHR and normalized them to LOINC codes. The laboratory tests (e.g., C-reactive protein) without system, quantity, time, type of scale, and type of method may map to multiple laboratory tests in EHR (e.g., C reactive protein, C reactive protein HS). Normalization of each laboratory test from EHR may map to multiple LOINC codes (e.g., LOINC codes, 1988-5, 14634-0, 11039-5, and 76485-2 for c reactive protein; LOINC codes, 30522-7 35648-5, 76486-0 and 59182-6 for c reactive protein HS). To simplify the mapping, we defined a set of rules to map each laboratory test in EHR to one LOINC code (e.g., 1988-5 for c reactive protein and 30522-7 for c reactive protein HS):

Rule 1. Mapping the most popular laboratory test in LOINC dictionary to the laboratory test in EHR, when the popularity rank is available in LOINC dictionary.

Rule 2. Mapping the laboratory test for serum and/or plasma samples in LOINC dictionary to the laboratory test in EHR, when the popularity rank is not available in LOINC dictionary.

Rule 3. When one-to-one mapping is not possible with Rule 1 and Rule 2, the test unit is applied to achieve the mapping.

Rule 4. When one-to-one mapping is not possible with Rule 1, Rule 2, and Rule 3, the unit gram is preferred than molar for mapping.

Rule 5. When one-to-one mapping is not possible with Rule 1, Rule 2, Rule 3, and Rule 4, the laboratory test without information about method is preferred for mapping.

A medication within the therapy domain can be mentioned with different synonyms across multiple EHR records. We normalized the medications by retrieving all the synonyms (i.e., generic name, brand name, and abbreviation) from Unified Medical Language System (UMLS) Metathesaurus[35] and RxNorm (https://mor.nlm.nih.gov/RxNav/). We observed that certain clinical characteristics from EHR within the diagnosis modifier domain were missing important information. For example, the breast cancer mentioned in EHR data from MSDW contains only Roman numbers (i.e. I, II, III, IV), not TNM stages. We normalized such clinical characteristics with the missing information (e.g., T1N0M0 = stage I) based on the NCCN guidelines. Additionally, "early stage, advanced stage, and metastasis" stages in eligibility criteria attributes were normalized the Roman numbers. Normalization of example eligibility criteria attributes and clinical characteristics is shown in Table 1. Three clinical domains, line of therapy, observation, and vital sign were not annotated and normalized prior to saving to the knowledge base.

Table 1. Example annotation and/or normalization of CT attributes and clinical characteristics

Domain	Source	Normalization

		Criteria attribute / Clinical characteristics	Concept from Standard Terminology / In-house Knowledge Base	Unique identifier	Standard Terminology / In-house Knowledge Base
Condition	Trial	Ulcerative Colitis	Ulcerative colitis	K51 (ICD-10)	ICD-10-CM ICD-9-CM
			Ulcerative (chronic) proctosigmoiditis	556.3 (ICD-9)	
			Left-sided ulcerative (chronic) colitis	556.5 (ICD-9)	
			Universal ulcerative (chronic) colitis	556.6 (ICD-9)	
			Other ulcerative colitis	556.8 (ICD-9)	
			Ulcerative (chronic) proctitis	556.2 (ICD-9)	
Procedure	EHR	Ileostomy	Laparoscopy, surgical; colectomy, total, abdominal, without proctectomy, with ileostomy or ileoproctostomy	44210	CPT-4
			Ileostomy or jejunostomy, non-tube	44310	
			Colectomy, total, abdominal, without proctectomy; with ileostomy or ileoproctostomy	44150	
			Colectomy, partial; with resection, with colostomy or ileostomy and creation of mucofistula	44144	
			Colectomy, partial; with coloproctostomy (low pelvic anastomosis)	44145	
			Laparoscopy, surgical; colectomy, total, abdominal, with proctectomy, with ileostomy	44212	
			Laparoscopy, surgical; colectomy, partial, with removal of terminal ileum with ileocolostomy	44205	
Laboratory Test	EHR	M protein UPEP	70663-M-SPIKE, %	33647-9	LOINC
			M-SPIKE MG/L	33358-3	
			M-SPIKE G/DL	33358-3	
			71280-M-SPIKE	33647-9	
Therapy	Trial	EGFR inhibitor	Panitumumab AMG-954 AMG954 Vectibix	Panitumumab	In-house Knowledge Base
			Rociletinib Xegafri AVL-301 AVL301 CO-1686 CO1686 CNX-419 CNX419	Rociletinib	

			Dacomitinib Vizimpro PF 00299804 PF-00299804 PF-299 PF299	Dacomitinib
			Cetuximab Erbitux BMS-564717 EMR-62202 IMC-C225, LY-2939777	Cetuximab
			Erlotinib Tarciva CP-358774 NSC 718781 OSI-774 R1415 R-1415 RG-1415 RG1415 Ro-50-8231 Ro50-8231	Erlotinib
			Gefitinib Iressa ZD-1839 ZD1839	Gefitinib
			Necitumumab Portrazza IMC-11F8 IMC11F8 LY-3012211 LY3012211	Necitumumab
			Osimertinib Tagrisso AZD-9291 AZD9291	Osimertinib
Biomarker	Trial	EGFR mutations sensitized to TKI	L858R in exon 21	L858R
			L861Q in exon 21	L861Q
			in-frame deletions in exon 19	in-frame deletions in exon 19
			deletions in exon 19 centered around four amino acids (LREA) at positions 747–750	deletions in 747–750
			deletion of leucine-747 to glutamic acid-749 (ΔLRE) in exon 19	deletion 747-749
			G719A in exon 18	G719A
			G719S in exon 18	G719S
			G719C in exon 18	G719C
			in-frame duplications and/or insertions in exon 20	in-frame duplications in exon 20

			in-frame duplications and/or insertions in exon 20	in-frame insertions in exon 20	
			S768I in exon 20	S768I	
			V765A in exon 20	V765A	
			T783A in exon 20	T783A	
Observation	-	-	-	-	-
Diagnosis modifier	Trials	T1N0M0	Stage I	Stage I	In-house Knowledge Base
Line of therapy	-	-	-	-	-
Vital sign	-	-	-	-	-
Demographic	Trials	Post-menopausal (<= 60 years, +LHRH agonist)	<= 60 years and taking Goserelin	<= 60 years Goserelin	In-house Knowledge Base
			<= 60 years and taking Leuprolide	<= 60 years Leuprolide	
			<= 60 years and taking Triptorelin	<= 60 years Triptorelin	
			<= 60 years and taking Histrelin	<= 60 years Histrelin	
			<= 60 years and taking Busereli	<= 60 years Buserelin	
			<= 60 years and taking Deslorelin	<= 60 years Deslorelin	

Quality assurance for semantic annotation and normalization

To address the challenge of exact matching between eligibility criteria attributes and clinical characteristics, we implemented two rules:

Rule A: We mapped eligibility criteria attributes to clinical characteristics at a higher or lower level within EHR or standard terminologies. For instance, the attribute "interstitial lung disease" could be mapped to more specific concepts in ICD-10-CM, such as acute respiratory distress syndrome, pulmonary edema, pulmonary eosinophilia, or other interstitial pulmonary diseases.

Rule B: We accounted for cases where an eligibility criteria attribute is part of a clinical characteristic, or standard terminologies include additional details. For example, the attribute "colectomy" could

correspond to a clinical characteristic like "colectomy/total/ostomy" or a standard terminology entry like "Colectomy, total; abdominal, without proctectomy; with ileostomy or ileoproctostomy."

We conducted quality assurance on a subset of annotated and normalized eligibility criteria attributes and EHR clinical characteristics in the domains of condition, procedure, laboratory test, and therapy. For the condition domain, we ensured the appropriateness of mapped ICD-9-CM or ICD-10-CM codes for eligibility criteria attributes, making necessary reassignments where needed. In procedure and laboratory test domains, we verified the correctness of mappings and LOINC codes between eligibility criteria attributes and EHR clinical characteristics. Corrections were made where needed, following defined rules. Regarding the therapy domain, we reviewed the medication list for completeness and accuracy. Adjustments were made, such as removing medications like Lapatinib that have dual roles as EGFR and HER2 inhibitors. In biomarkers, observations, and diagnosis modifiers domains, we reviewed the annotation completeness and correctness for each attribute, making updates based on careful examination. For instance, additional single-point substitutions in EGFR were added to the mutation list for EGFR mutations sensitized to tyrosine kinase in non-small cell lung cancer. Through the implementation of these rules and thorough reviews, we ensured the quality and accuracy of the annotated and normalized eligibility criteria attributes and EHR clinical characteristics, enhancing the reliability of the data for further analysis and research.

Clinical phenotyping knowledge base

The annotated and normalized eligibility criteria attributes were indexed and stored in a Redshift database. The normalized clinical characteristics from EHR were also stored in Redshift database as reference tables. The indexed eligibility criteria attributes and reference tables together form the knowledge base for clinical phenotyping.

Validation and Quality Control*:* We carried out two types of validations. First, two curators (YM and KL) evaluated a subset of the annotation from the Redshift database. The inter-rater's agreement was measured by Cohen's Kappa coefficient [31]. Second, we retrieved a subset of randomly selected eligibility criteria attributes that were annotated by the experts and evaluated it using a gold standard, which was generated from EHR. The gold standard includes EHR data for a subset of randomly selected patients with the desired clinical characteristics (e.g., patient age at the time of phenotyping, the diagnosed conditions before the phenotyping was performed) to be reviewed. We mapped the eligibility criteria attributes of the eligibility criteria of clinical trials to the clinical characteristics of EHR using patient ID and date. We reported the performance using the standard metrics, precision, recall, and F1-score.

Results

Annotated and normalized attributes

We extracted 640 unique attributes with values from 3,475 clinical trials (the whole list is provided in Multimedia Appendix 2) and grouped them under 10 clinical domains (Table 2). 367 out of 640 attributes (57.34%), belonging to seven clinical domains, condition, procedure, laboratory test, therapy, biomarker, observation, and diagnosis modifier, were annotated and normalized prior to storing to the Redshift database. Among the 363 annotated and normalized attributes, 174 attributes (47.41%) were normalized using the standard terminology and 193 attributes (52.59%) were normalized using the concepts from the reference tables. While 72 attributes under laboratory tests were normalized using the standard terminology alone, two attributes under biomarker and one attribute under observation were normalized using the reference tables only. Normalization of attributes under therapy and diagnosis modifier were mainly achieved with the reference tables. In the therapy domain, three attributes were normalized using standard terminology and 163 attributes were normalized using reference tables. In the diagnosis modifier domain, seven attributes were normalized using standard terminology, and 18 attributes were normalized using reference tables. Our results show that EHR includes several attributes that are not in standard terminologies such as CPT, ICD-9-CM, and ICD-10-CM. The gap between EHR and the standard terminologies was filled with our reference tables. We did not annotate or normalize 273 attributes (42.66%) because 133 attributes (48.72%) did not require annotation or normalization (e.g., age), 140 attributes (51.28%) were difficult to achieve mapping to EHR (e.g., disease status “in remission/respond” and "unresolved toxicity from the prior treatment”).

Table 2 Annotated and normalized attributes of the eligibility criteria of clinical trials and the clinical characteristic of EHR

Clinical Domain	Attributes (all)	Attributes (Annotated and normalized)	Attributes Normalization	
			Standard terminology	Reference tables
Condition	133 (20.78%)	84	79	5
Procedure	22 (3.44%)	17	13	4
Laboratory Test	81 (12.66%)	72	72	0
Therapy	214 (33.44%)	166	3	163
Biomarker	59 (9.22%)	2	0	2
Observation	11 (1.72%)	1	0	1
Diagnosis Modifier	68 (10.63%)	25	7	18
Line of Therapy	5 (0.78%)	0	0	0
Vital Sign	27 (4.22%)	0	0	0
Demographic	20 (3.13%)	0	0	0

Attributes Distribution

The majority of the annotated and normalized eligibility criteria attributes are from three domains: condition, laboratory test, and therapy. These attributes are dominantly found in clinical trials for non-small cell lung cancer, prostate cancer, and breast cancer (Fig. 2). Conversely, the unannotated and unnormalized attributes belong to seven groups: demographic, disease index, line of therapy, neoadjuvant treatment, radiotherapy, vital, and other (See Multimedia Appendix 3 for details). The annotated and normalized attributes of the eligibility criteria of clinical trial belong to 28 attribute groups (Fig. 3A). Among, four groups, test, targeted therapy, hormone therapy, and medication, were frequently mentioned in the eligibility criteria of the clinical trials (i.e., 58.31% of all annotated clinical trial attributes).

Fig. 2. Clinical phenotypes in different clinical domains. (A) Distribution of annotated/normalized attribute across different clinical domains. (B) Distribution of annotated/normalized attributes of each disease across different clinical domains.

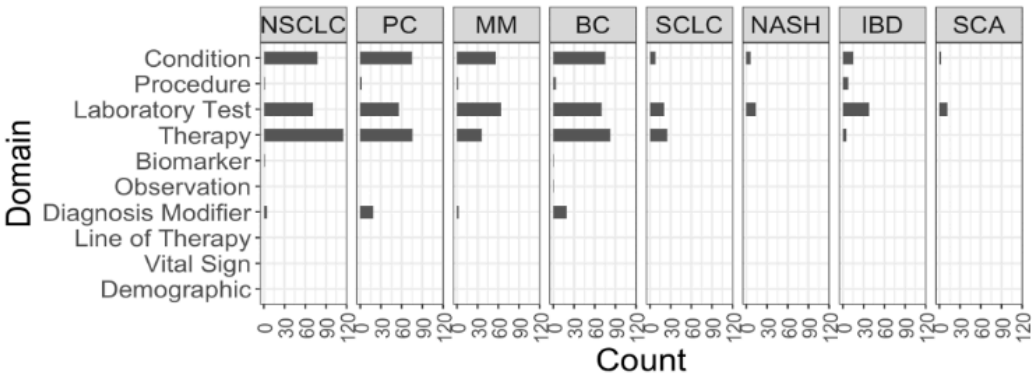
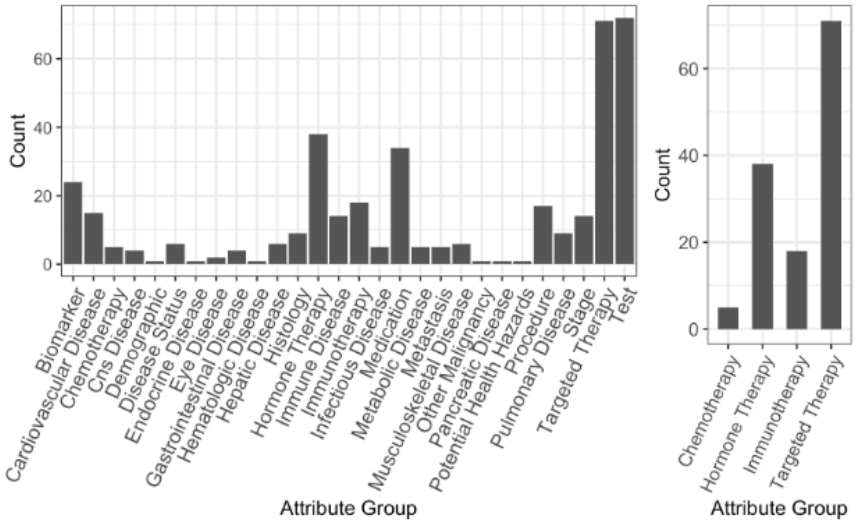


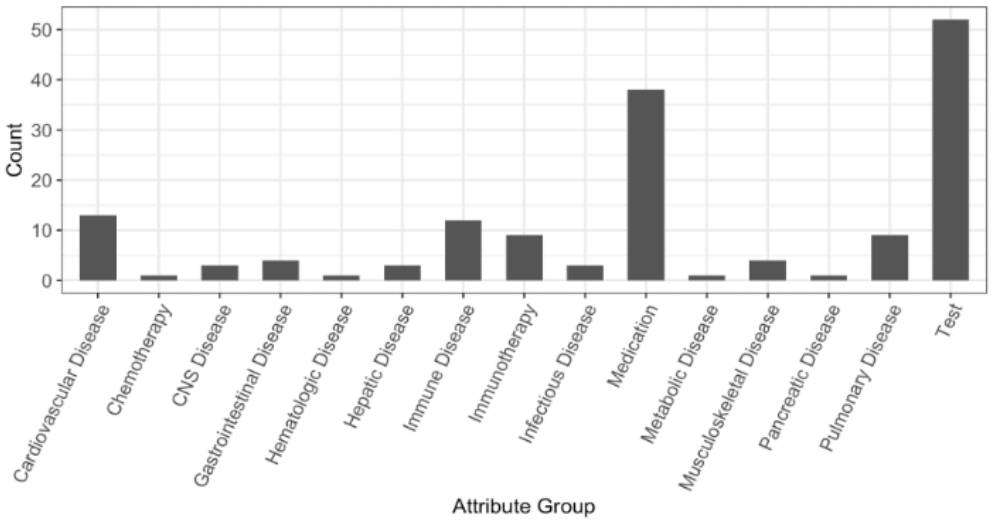
Fig 3. Clinical phenotypes in different attribute groups. (A) Distribution of annotated/normalized attributes across different attribute groups. (B) Distribution of annotated/normalized attributes across different modalities in clinical trials of cancer treatment.



Three of the top four eligibility criteria attribute groups, medication, targeted therapy and hormone therapy are related to the treatments: (i) treatments for comorbidities that are to be excluded, (ii) treatments that will interfere with the clinical trial, or (iii) treatments related to the diseases under study. The eligibility criteria attribute groups within the therapy domain represent the cancer therapies that comprise of regimen or medications used in cancer treatment. The eligibility criteria attribute group, medication, includes drugs for treating cancer. The clinical trial attribute groups, targeted therapy and hormone therapy, are from the eligibility criteria of cancer clinical trials. The drugs used in cancer treatment were regrouped into four attribute groups namely chemotherapy, targeted therapy, immunotherapy, and hormone therapy. Among these attribute groups, 51.91% (73/1231) belong to targeted therapy, 23.66% (31/131) belong to hormone therapy, 13.74% (18/131) belong to immunotherapy, and only 3.82% (5/131) belong to chemotherapy (Fig. 3B). The targeted therapy and hormone therapy are the most frequently mentioned treatment options for cancer.

We observed a set of commonly used attributes in the eligibility criteria of clinical trials related to cancer (Fig. 4). These attributes describe the conditions (e.g., cardiovascular disease), treatments for these condition (i.e. medication), previous line of therapy for cancer (e.g. chemotherapy), and laboratory test in the eligibility criteria (blood, liver, and kidney function tests). These attributes may be considered when making decision for eligibility criteria of cancer clinical trials.

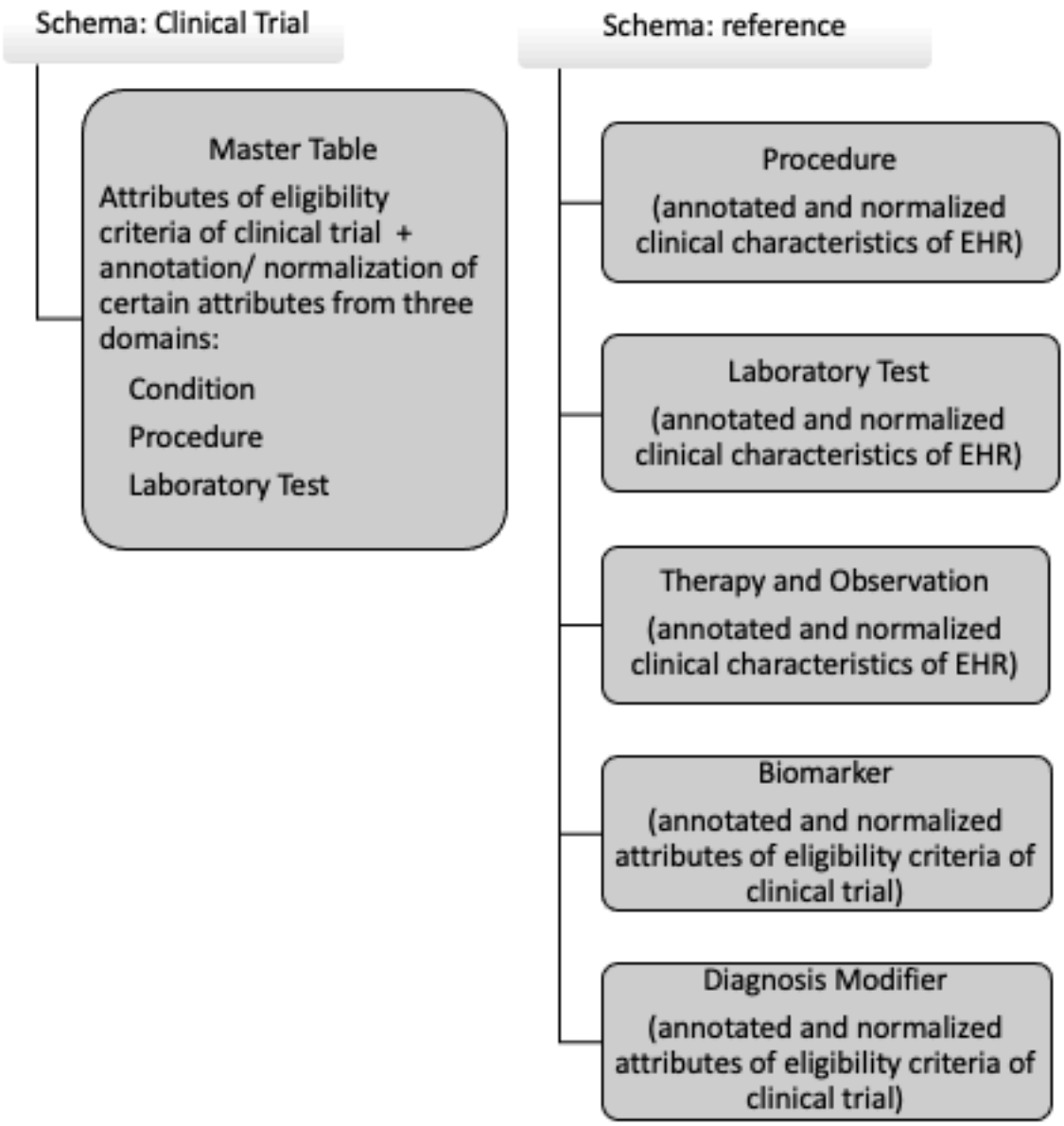
Fig. 4. Common clinical phenotypes in clinical trials of cancer studies.



Knowledge Base

Our knowledge base consists of two schema, clinical trials, and reference. Fig. 5 illustrates the schema and tables included in the knowledge base. The annotated and normalized attributes of eligibility criteria of clinical trials from three domains, condition, procedure, and laboratory test were stored together in a master table under schema for clinical trial. The annotated and normalized clinical characteristics of EHR such as procedure, lab test, biomarker, and diagnosis modifier were stored in separate tables under schema for reference. The annotated and normalized clinical characteristics of EHR from two domains, therapy and observation were stored in one table. Each record can be queried using the attribute ID or attribute name.

Fig. 5. Knowledge base for annotated and normalized eligibility criteria attributes and normalized clinical characteristics of EHR.



Evaluation of semantic annotation and normalization: Table 3 shows the outcome of the quality control performed on a randomly selected subset of annotated and normalized clinical trial attributes of the eligibility criteria of clinical trials and clinical characteristics of EHR within five domains, condition, procedure, laboratory test, therapy, and diagnosis modifier.

Table 3. Quality control

Domain	Attributes Reviewed	Attributes Modified
Condition	33	10 (30%)
Procedure	5	4 (80%)
Laboratory test	11	0
Therapy	15	1 (6.67%)
Diagnosis Modifier	3	0

Evaluation of Clinical phenotyping knowledge base: The inter-rater agreement on the annotation of a random subset (89 out of 260 clinical trial attributes) of the knowledge base measured by Cohen's Kappa coefficient is 0.82 (p = 0). The accuracy of patient matching is and the was 0.94 F1-score (Table 4). The knowledge base was also successfully applied to EHR data from other institutes (data not shown) for patient pre-screening, suggesting its generalization capability.

Table 4. Evaluation of clinical attributes

Domain	Attribute group	Attribute name	Attribute value	Precision	Recall	F1	specificity
Condition	Other malignancy	Other primary Malignancy	<= 5Y	0.83	1	0.91	0.8
	Cardiovascular Disease	Congestive Heart Failure	Yes	1	1	1	1
	Histology	Squamous NSCLC	Yes	1	1	1	1
Procedure	Procedure	Organ/Tissue Transplantation	Yes	1	1	1	1

Laboratory test	Test	Platelets	>=75000	1	1	1	1
Therapy	Immunotherapy	PD-1 Ab	Pembrolizumab	1	1	1	1
Biomarker	Biomarker	PD-1/PD-L1 Positive	Yes	1	0.75	0.86	1
Diagnosis modifier	Stage	Stage Groups	Extensive stage	0.8	1	0.89	0.83
Line of Therapy	Line of Therapy	Prior LOT	1	0.7	1	0.82	0
Vital Sign	Vital	ECOG	0	1	1	1	1

Discussion

In this study, we built an intermediate representation of annotated and normalized attributes from the eligibility criteria of clinical trials and the clinical characteristics found in EHR for clinical phenotyping. These annotated and normalized attributes facilitate the usability and interoperability of EHR data across multiple healthcare observational databases, making it easier to identify potentially eligible patients for clinical trials. The majority (87.74%) of the annotation and normalization work focused on three domains: condition, laboratory test, and therapy. These three domains were consistently mentioned in the eligibility criteria of clinical trials across all the diseases analyzed. Therefore, the standardization of EHR data related to therapy, condition, and laboratory test through standard terminology was prioritized to facilitate the development of an intermediate representation for eligibility criteria clinical phenotyping.

In cancer clinical trials, targeted therapy and hormone therapy were more frequently mentioned than other types of therapy or modality. Immunotherapy had a smaller number of attributes compared to hormone therapy (47.37%) and targeted therapy (25.35%), but a greater number of attributes than chemotherapy (~ 4%) (Figure 3). The last few decades have witnessed significant advancements in our understanding of molecular pathogenesis and the identification of novel disease-driven genetic disorders. These discoveries have led to the introduction of numerous targeted therapies, hormone therapy, and immunotherapy were introduced in cancer treatment. Currently, many of these therapies are being investigated in clinical trials and often aim to recruit subjects with relevant genetic alterations. Due to limited biomarker data in the current EHR database, a lower number of eligibility criteria attributes from the biomarker domain was annotated and normalized (0.31 %) in this study. Expanding biomarker measurements in real-world would be beneficial for the advancing precision medicine.

We phenotyped 92.37% of eligibility criteria attributes (339 out of 367) in the domain of condition, procedure, laboratory test, and therapy. However, certain attributes including (i)

CDAI (CD activity index), a diagnosis modifier attribute, (ii) fecal microbial transplantation, a procedure attribute, and (iii) NaPi2b targeted therapy, a therapy attribute, were not be phenotyped due to unavailability of data in the structured EHR data in MSDW and VieCure. In the future work, an alternative approach can be explored by leveraging data from the clinical notes for phenotyping. In our previous work (https://preprints.jmir.org/preprint/50800), we implemented advanced deep-learning NLP techniques using Conditional Random Fields (CRF) and Bi-directional Long Short Term Memory to extract attributes from clinical trial eligibility criteria. This pipeline can be further expanded to process clinical notes, enabling the automated phenotyping of attributes in clinical trial eligibility criteria from huge text-based data.

Limitation

Our study has several limitations. Firstly, limited biomarker data available in the EHR database. Expanding biomarker measurements in real-world EHR data could improve the precision of phenotyping for clinical trials. Secondly, unavailability of certain eligibility criteria eligibility. Exploring alternative approaches, such as leveraging data from clinical notes, may help address this issue in future work. Thirdly, our normalization approach was carried out manually. The study acknowledges that using the billing code such as CPT for laboratory test, and the standard encoding information such as NDC (National Drug Code Dictionary) code for medications could automate the normalization process and accelerated the normalization of clinical characteristics of her.

Additionally, leveraging the unique concept identifier (CUI) from UMLS Metathesaurus generated during data extraction using NLP can aid in automating the normalization of eligibility criteria attributes.

Moreover, our study focused on only one arm of clinical trials for clinical phenotyping. Future work aims to include attributes from every arm of the clinical trials to enhance the comprehensiveness of the analysis and further enrich the knowledge base.

Conclusions

We developed a clinical trial phenotyping pipeline and knowledge base that maps clinical trial attributes to EHR clinical characteristics. This enables automated cohort selection for clinical trials and exhibits generalization across different institutes. Our approach complements standard terminologies, enhancing the normalization of clinical attributes and facilitating efficient patient matching for research.

Conflict of interest

YM, KL, KR, ZL, TJ, MM, MH, TW, LA, CE, and XW are employees of Sema4. WO and ES are employees of the Icahn School of Medicine at Mount Sinai. All Authors declare no other competing financial or non-financial interests.

Authors' contribution

Y. Mai, K. Lee, K. Raja and X. Wang designed the study and wrote the manuscript. Y. Mai and K. Lee annotated clinical trial eligibility criteria, patient notes, and published knowledge bases. Y. Mai, Z. Liu, M. Ma, and T. Wang were involved in the data analysis. M.K. Higashi, T. Jun, L. Ai, E. Calay, W. Oh, E. Schadt, X. Wang discussed the project and reviewed the manuscript.

Reference

1 Ulrich, C. M. *et al.* RTOG physician and research associate attitudes, beliefs and practices regarding clinical trials: implications for improving patient recruitment. *Contemp Clin Trials* **31**, 221-228, doi:10.1016/j.cct.2010.03.002 (2010).

2 Unger JM, C. E., Tai E, Bleyer A. The Role of Clinical Trial Participation in Cancer Research: Barriers, Evidence, and Strategies. *Am Soc Clin Oncol Educ Book* **35**, 185-198 (2016).

3 Augustine, E. F., Adams, H. R. & Mink, J. W. Clinical trials in rare disease: challenges and opportunities. *J Child Neurol* **28**, 1142-1150, doi:10.1177/0883073813495959 (2013).

4 Rothwell, P. M. External validity of randomised controlled trials: "to whom do the results of this trial apply?". *Lancet* **365**, 82-93, doi:10.1016/S0140-6736(04)17670-8 (2005).

5 Van Spall, H. G., Toren, A., Kiss, A. & Fowler, R. A. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* **297**, 1233-1240, doi:10.1001/jama.297.11.1233 (2007).

6 (The National Academies Press, Washington, DC, 2010).

7 Alexander, M. *et al.* Evaluation of an artificial intelligence clinical trial matching system in Australian lung cancer patients. *JAMIA Open* **3**, 209-215, doi:10.1093/jamiaopen/ooaa002 (2020).

8 Angus, D. C. Fusing Randomized Trials With Big Data: The Key to Self-learning Health Care Systems? *JAMA* **314**, 767-768, doi:10.1001/jama.2015.7762 (2015).

9 Beck, J. T. *et al.* Artificial Intelligence Tool for Optimizing Eligibility Screening for Clinical Trials in a Large Community Cancer Center. *JCO Clin Cancer Inform* **4**, 50-59, doi:10.1200/CCI.19.00079 (2020).

10 Meystre, S. M., Heider, P. M., Kim, Y., Aruch, D. B. & Britten, C. D. Automatic trial eligibility surveillance based on unstructured clinical data. *Int J Med Inform* **129**, 13-19, doi:10.1016/j.ijmedinf.2019.05.018 (2019).

11 Ni, Y. *et al.* Increasing the efficiency of trial-patient matching: automated clinical trial eligibility pre-screening for pediatric oncology patients. *BMC Med Inform Decis Mak* **15**, 28, doi:10.1186/s12911-015-0149-3 (2015).

12 Shivade, C. *et al.* A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc* **21**, 221-230, doi:10.1136/amiajnl-2013-001935 (2014).

13 Richesson, R. L., Sun, J., Pathak, J., Kho, A. N. & Denny, J. C. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artif Intell Med* **71**, 57-61, doi:10.1016/j.artmed.2016.05.005 (2016).

14 Pathak, J., Kho, A. N. & Denny, J. C. Electronic health records-driven phenotyping: challenges, recent advances, and perspectives. *J Am Med Inform Assoc* **20**, e206-211, doi:10.1136/amiajnl-2013-002428 (2013).

15 Ping Wang, T. S., Chandan K. Reddy. Text-to-SQL Generation for Question Answering on Electronic Medical Records. (2020).

16 Xiaojing Yu, T. C., Zhengjie Yu, Huiyu Li, Yang Yang, Xiaoqian Jiang, Anxiao Jiang. in *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC)* 5829–5837 (2020).

17 Yuan, C. *et al.* Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc* **26**, 294-305, doi:10.1093/jamia/ocy178 (2019).

18 Bodenreider, O. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearb Med Inform*, 67-79 (2008).

19 Chondrogiannis, E. *et al.* A novel semantic representation for eligibility criteria in clinical trials. *J Biomed Inform* **69**, 10-23, doi:10.1016/j.jbi.2017.03.013 (2017).

20 Hassanzadeh H, K. S., Nguyen A. Matching patients to clinical trials using semantically enriched document representation. *J. Biomed. Informatics* **105** (2020).

21 Hersh, W. R. & Greenes, R. A. SAPHIRE--an information retrieval system featuring concept matching, automatic indexing, probabilistic retrieval, and hierarchical relationships. *Comput Biomed Res* **23**, 410-425, doi:10.1016/0010-4809(90)90031-7 (1990).

22 Hongfang Liu, S. J. B., Sunghwan Sohn, Sean Murphy, Kavishwar B Wagholikar, Siddhartha R Jonnalagadda, K E Ravikumar, Stephen T Wu, Iftikhar J Kullo, Christopher G Chute. in *Proc. AMIA Jt. Summits Transl. Sci* 149–153 (2013).

23 Hripcsak G, C. P., Pryor TA, Haug P, Wigertz OB, Van der lei J. in *Proc Annu Symp Comput Appl Med Care* 200-204 (Washington, DC, 1990).

24 Richesson, R. L. *et al.* Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc* **20**, e226-231, doi:10.1136/amiajnl-2013-001926 (2013).

25 Weng, C., Tu, S. W., Sim, I. & Richesson, R. Formal representation of eligibility criteria: a literature review. *J Biomed Inform* **43**, 451-467, doi:10.1016/j.jbi.2009.12.004 (2010).

26 Lonsdale, D. W., Tustison, C., Parker, C. G., and Embley, D. W. Assessing clinical trial eligibility with logic expression queries. *Data & Knowledge Engineering* **66**, 3–17 (2008).

27 Sordo, M., Boxwala, A. A., Ogunyemi, O. & Greenes, R. A. Description and status update on GELLO: a proposed standardized object-oriented expression language for clinical decision support. *Stud Health Technol Inform* **107**, 164-168 (2004).

28 Delaney, R. B. A. T. S. M. B. C. An Eligibility Criteria Query Language for Heterogeneous Data Warehouses. *Methods Inf Med* **54**, 41-44 (2015).

29 Lindsay J, D. V. F. C., Zwiesler Z. Algorithmic matching of genomic profiles to precision cancer medicine clinical trials at DFCI. *Journal of Clinical Oncology* **35**, 6620-6620 (2017).

30 Lindsay J, D. V. F. C., Zwiesler Z. MatchMiner: An open source computational platform for real-time matching of cancer patients to precision medicine clinical trials using genomic and clinical criteria. . (2017). <https://doi.org/10.1101/199489>.

31 Tu, S. W. *et al.* A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform* **44**, 239-250, doi:10.1016/j.jbi.2010.09.007 (2011).

32 Weng, C. *et al.* EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* **18 Suppl 1**, i116-124, doi:10.1136/amiajnl-2011-000321 (2011).

33 Kang, T. *et al.* EliIE: An open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc* **24**, 1062-1071, doi:10.1093/jamia/ocx019 (2017).

34 Levy-Fix, G., Yaman, A., & Weng, C. Structuring Clinical Trial Eligibility Criteria with the Common Data Model. (2014).

35 Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32**, D267-270, doi:10.1093/nar/gkh061 (2004).