The Data Gap in the EHR for Clinical Research Eligibility Screening

Alex Butler, BA[1], Wei Wei, PhD[1], Chi Yuan, MS[1], Tian Kang, MA[1], Yuqi Si, MS[2], Chunhua Weng, PhD[1]

[1]Department of Biomedical Informatics, Columbia University, New York City, New York; [2]The University of Texas Health Science Center at Houston, Houston, Texas

Abstract

Much effort has been devoted to leverage EHR data for matching patients into clinical trials. However, EHRs may not contain all important data elements for clinical research eligibility screening. To better design research-friendly EHRs, an important step is to identify data elements frequently used for eligibility screening but not yet available in EHRs. This study fills this knowledge gap. Using the Alzheimer's disease domain as an example, we performed text mining on the eligibility criteria text in Clinicaltrials.gov to identify frequently used eligibility criteria concepts. We compared them to the EHR data elements of a cohort of Alzheimer's Disease patients to assess the data gap by using the OMOP Common Data Model to standardize the representations for both criteria concepts and EHR data elements. We identified the most common SNOMED CT concepts used in Alzheimer's Disease trials, and found 40% of common eligibility criteria concepts were not even defined in the concept space in the EHR dataset for a cohort of Alzheimer's Disease patients, indicating a significant data gap may impede EHR-based eligibility screening. The results of this study can be useful for designing targeted research data collection forms to help fill the data gap in the EHR.

Introduction

Randomized clinical trials (RCTs) are the well-regarded gold standard for generating high-quality medical evidence[1]. The success of RCTs depends on successful enrollment[1,2], which remains the No.1 barrier to RCTs. According to the recent statistics, only 2-4% of adult patients with cancer participate in RCTs, and this number remained unchanged since 1994[2,3]. Inefficient or unrepresentative participant recruitment can cause study delays, increase costs, weaken the statistical power of analysis, and finally, may lead to failed clinical trials[4].

A major bottleneck step in RCT recruitment is eligibility screening[2]. However, conventional methods for eligibility screening involves laborious manual review of the syntactic rules and semantic concepts in eligibility criteria and clinical data sources[5,6]. This process is not only time-consuming, but also expensive: the cost of eligibility screening is usually not compensated through contracts supporting CTs, and the expense can go up to $336.48 per participant[2].

Much effort[4,7,8] has been made to advance automated identification of eligible patients in the biomedical informatics research community. In the meantime, Electronic Health Record (EHR) data have been recognized as an important clinical data source and were adopted in multiple automated identification methods[7-10]. EHR-based automated approaches have been reported to reduce workload by up to 90%[7] and almost reached the theoretical maximum area under ROC curve[8].

A concern of EHR-based eligibility screening is that EHRs may not contain all important data frequently used for eligibility screening since EHRs are designed for patient care rather than clinical research. Our previous study in cancer trial eligibility criteria showed that a lot of eligibility criteria used in cancer trials are not present in EHR data so that clinical research coordinators creatively invented a list of "major eligibility criteria" for patient screening to optimize the efficiency of patient screening[11]. A recent study by Köpcke et al. showed that on average 55% of eligibility criteria required data elements are present in EHR. However, there are three major limitations of their study: (1) only numeric and structured data elements in EHRs like checkboxes and dropdown menus were included in analyses so that EHR narratives were excluded; (2) EHR data from five participating hospitals were not harmonized using any common data model, resulting in unaccounted overlaps or inconsistency among available EHR data elements across sites; (3) the whole process was manual so that patient characteristics (i.e., clinical entities) were manually identified from free-text eligibility criteria followed by assignment of semantic categories, which were again manually mapped to EHR data elements, making their method not scalable.

This study presented here shares the same goal of Köpcke's study but contributes a novel scalable data-driven approach by leveraging the public clinical trial summary text and the publicly available synthetic clinical data. Next we will describe our methodology details and results as well as implications.

Methods

To overcome the limitations of Köpcke's study, we extracted common data elements from free-text eligibility criteria[12,13] for Alzheimer's disease (AD) and represented both EHR data elements and eligibility criteria concepts using The Observational Medical Outcomes Partnership (OMOP)[14] Common Data Model (CDM) supported by the Observational Health Data Sciences and Informatics (OHDSI)[15] consortium (**Figure 1**).

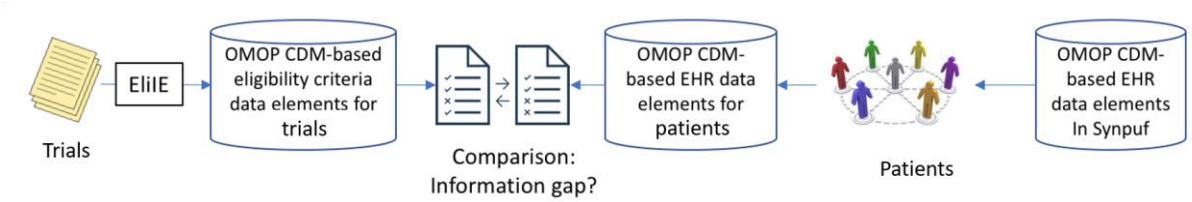

Figure 1. Overview of the study design for comparing OMOP CDM-based criteria and EHR data for AD trials

The OMOP CDM has been adopted by active scientific consortiums such as OHDSI[15] and eMERGE[16], and has included about 1.26 billion patients as of October 2017. The OMOP CDM-standardized EHR ensures the semantic interoperability of EHR data from multiple participating sites. The sheer number of patients will allow large sample sizes and likely lead to more generalizable study results. Free-text eligibility criteria were automatically processed using Eligibility Criteria Information Extraction (EliIE)[12], an open-source information extraction system for structuring eligibility criteria according to the OMOP CDM, and then extracted information (e.g., clinical entities) was stored in a relational database[13]. The fully automated eligibility criteria processing techniques make our method highly scalable and improve the efficiency of large-scale studies.

As the first step for methodology illustration, we used eligibility criteria from 1,587 clinical trials for AD and a de-identified EHR dataset, Synthesized Public Use File (SynPUF) 1%, to study the data gap. The publicly available SynPUF 1% dataset, which includes a set of over 116,350 patients' de-identified EHR structured data points, served as the clinical data source. We mapped clinical entities in eligibility criteria to The Systematized NOMenclature of MEDicine – Clinical Terms (SNOMED CT)[17] terms (hereafter referred to as "variables"), merged relevant variables and created a list of unique common variables. SNOMED CT was chosen as the ultimate clinical database in this analysis because it has been preferred as the encoding terminology for clinical concepts by researchers on various other projects.[18] We picked a subjective threshold of "being present in at least 15 trials" to select common variables, visualized the relations among the variables and their parents, and analyzed the prevalence of the variables in an EHR dataset. For the purposes of this analysis, we focused on the 19,570 patients who had a previous diagnosis of Alzheimer's Disease within the SynPUF 1% dataset (hereafter referred to as "the EHR dataset") as the clinical data source in this study. OHDSI ATLAS, a web-based open source software available at http://www.ohdsi.org/web/atlas for scientific analyses of observational data was adopted to identify qualified patient records from the EHR. The details of are provided below (Figure 2).

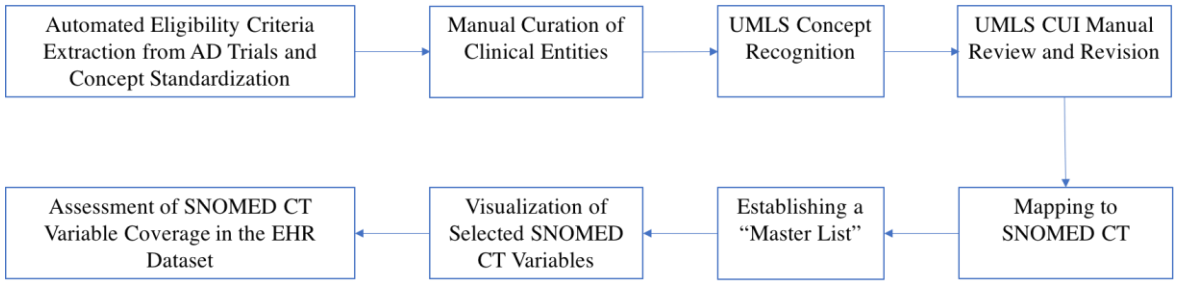

Figure 2.The eight-step workflow of this study.

Step 1. Automated Eligibility Criteria Extraction from AD Trials and Concept Standardization

Free-text eligibility criteria were downloaded from The ClinicalTrials.gov, reformatted using the previously published open-source EliIE[12] system, and stored in a public relational database (https://github.com/Yuqi92/DBMS_EC)[13]. All the eligibility criteria of 1,587 AD trials (collected until September 2016) were represented using the OMOP CDM v5.0 model, which allows focusing on four classes of entities: ***condition***, ***observation***, ***drug/substance***, and ***procedure***

or device. A total of 9,261 unique clinical entities were identified[13] from all of the eligibility criteria. For analysis, corresponding modifiers (e.g., qualifier, measurement) and inclusion/exclusion status were attached to each entity.

Step 2. Manual Curation of Clinical Entities

A manual review of unique clinical entities was performed by a medical student (AB). Modifications were made to produce a simplified list of clinical entities (e.g., AD was used to refer to Alzheimer's disease). To identify the relevant entities, all entities were sorted alphabetically, so word-similar entity comparison was possible as has been done algorithmically by Varghese and Dugas[19]. All reasons for modification were captured and can serve as evidence in the future for eligibility criteria terminology guidelines.

Step 3. UMLS Concept Recognition

The clinical entities in the simplified list were mapped to the Unified Medical Language System (UMLS) Metathesaurus[20], which was chosen because it is the largest thesaurus in the biomedical domain.[21] This mapping was performed via a widely adopted NLP system developed by the National Library of Medicine, MetaMap[22]. MetaMap was chosen over other NLP systems because of its widespread adoption, easy learning curve and batch request functionality, which allowed large blocks of text to be analyzed simultaneously. For clarity, all phrases contained in the original entity list will be referred to as "entities" and all terms found in the UMLS Metathesaurus will be referred to as "concepts.". The configuration of MetaMap query options were as below:

-JSONf 2 (formatted JSON output),
-g (Allow Concept Gaps),
-z (Term Processing),
-Q 4 (Composite Phrases),
-y (Use Word Sense Disambiguation),
-E (Indicate Citation End; required for batch scheduler)

Figure 3 illustrates this concept recognition process. When multiple phrases contain one or more concepts in a query, the term with the highest MetaMap score was retrieved. In the case that multiple phrases containing 1 or more concept were returned with identical MetaMap scores, the phrase with the lowest level of clinical specificity was chosen to not exclude any concepts. Review of the simplified entity list found numerous multi-term entities, so single term retrieval was not performed.

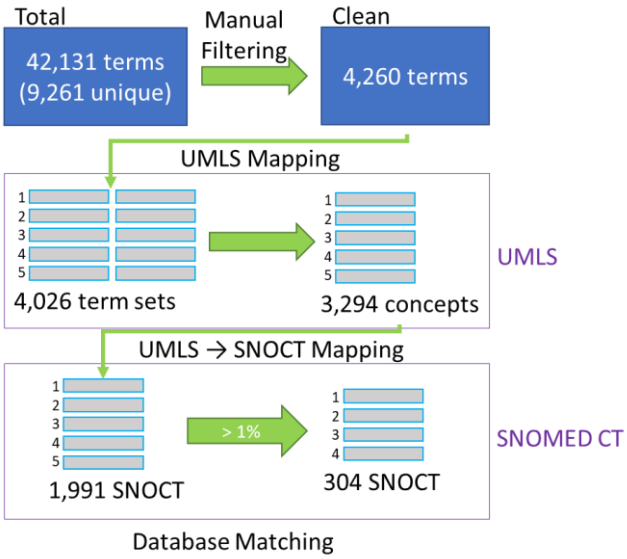

Figure 3. The process of deriving SNOMED CT terms from clinical trial eligibility criteria. 42,131 clinical entities were extracted from the eligibility criteria of 1,587 clinical trials. A simplified list of 4,260 clinical entities was generated following manual review and filtration, and this list was mapped first to 3,294 UMLS concepts, and then to 1,991 SNOMED CT variables, of which 304 variables occur in more than 1% of all trials (i.e., 15 trials).

Step 4. UMLS CUI Manual Review and Revision

There were a number of data quality issues identified when performing concept extraction. A total of 3,610 manual edits were made to the "master list" for clinical entities as tracked by our computer with the six main types, including typos, plural, trimmed, other formatting reason, simplification, and multi-term (Table 1). Therefore, the identified UMLS concepts and associated Concept Unique Identifiers (CUIs) were manually reviewed and corrected by a medical student (AB). The corrections were performed for two primary reasons: (1) simple corrections which are applied when the CUI of a concept is replaced by a more appropriate CUI, and (2) type corrections which are applied when the CUI of a concept is replaced by a CUI of a more appropriate type according to UMLS coding.

Step 5. Mapping to SNOMED CT

For every UMLS concept, its corresponding term in SNOMED CT was identified. Due to the design of UMLS Metathesaurus as a hub for numerous terminologies, the SNOMED CT variables associated with the UMLS concepts were used when such variables were possible. In the case that no SNOMED CT variable was found, a manual search of the SNOMED CT terminology was conducted to identify the closest available match (Figure 3). Manual modifications were also performed for SNOMED CT types which were inappropriate for use in eligibility screening. For example, "alanine aminotransferase (substance)" was changed to "alanine aminotransferase measurement (procedure)."

Table 1. Manual revision of clinical entities.

Types of Revision	Example	Times
Formatting; Typo	delerium -> delirium	207
Formatting; Plural	cancers -> cancer	253
Formatting; removal of non-informative words	heart rate measurement -> heart rate	364
Formatting; removal of abbreviations	absolute neutrophil count (ANC) -> absolute neutrophil count	1768
Simplification	asthmatic conditions -> asthma	573
Breaking down long phrases to logically-connected single phrases	basal or squamous cell carcinoma -> basal cell carcinoma or squamous cell carcinoma	445
Total		**3610**

Step 6. Establishing a "Master List"

Trial occurrences were tracked for each clinical entity and carried through to mapped SNOMED CT variables to calculate an overall trial frequency. SNOMED CT variables chosen for the "master list" were found in at least 1% of all trials, meaning they were used as an eligibility criterion in at least 15 trials.

Step 7. Visualization of Selected SNOMED CT Variables

Since SNOMED CT maintains a hierarchical structure, the parents of all variables present in the "master list" were captured. All of the "master list" variables, their parent variables, and the "is-a" hierarchical relations were stored in JSON files and visualized using a modified d3js package. Also, the trial frequency for each variable was also obtained and stored within the corresponding JSON file. Of note, every parent of a "master list" variable was considered to have the same trial frequency as its child.

Step 8. Assessment of SNOMED CT Variable Coverage in the EHR Dataset

The SNOMED CT ID associated with each SNOMED CT variable in the "master list" was queried in ATLAS and the record count (RC) and descendant record count (DRC) were returned. RC indicates the number of times a specific variable is found in the EHR dataset, and DRC indicates the number of times a specific variable and its descendants are found in the dataset. SNOMED CT variables were further classified into five sets:

(1) categorical variables (e.g., the presence of Parkinson's Disease) that are available in EHR
(2) continuous variables (e.g., age) that are available in EHR
(3) variable not found in EHR, but can be derived from the existing EHR variable, such as "chronological age" can be derived from variable "date of birth"

(4) variables not available in EHR, but the data could be collected from a patient without medical training, such as questions in Mini-Mental Status Exam
(5) variables not available in EHR, and the information could not be provided by a patient without medical training, such as "General Metabolic Function"
(6) variables not found in EHR, and not relevant for eligibility screening, such as "Psychiatric"

Results

The 42,131 entities identified in clinical trial eligibility criteria contained 9,261 unique entities, 1,930 of which corresponded to medication information which were not included in this analysis. Manual review of the remaining 7,331 unique non-medication entities simplified the list to 4,260 entities. To reach this simplified list, 3,610 manual changes were made. 2,591 changes were made for formatting reasons (e.g. AD, AD Disease -> Alzheimer's Disease), 574 changes were made for simplification reasons (e.g. asthmatic conditions, adult asthma -> asthma) and 445 changes were made for 'multi-term' entities (e.g. basal or squamous cell carcinoma -> basal cell carcinoma or squamous cell carcinoma). A total of 4,260 unique clinical concepts were mapped to UMLS concepts via MetaMap, resulting in 4,026 unique MetaMap term sets (e.g. basal cell carcinoma or squamous cell carcinoma is a single 'term set' as the phrase was extracted from an eligibility criterion, but each underlined section is handled as a separate UMLS concept). A total of 111 manual searches were performed, including 66 searches for multi-term clinical entities, one for a typo in the entity, and 44 for inaccurate MetaMap mapping as assessed by the medical student (AB). After sorting, the final UMLS concept list was composed of 3,294 unique concepts. Of note, it was observed on manual review that many of the lab tests being used for eligibility assessment were found to be of UMLS type "Amino Acid, Peptide, or Protein" so all concepts of this type were re-queried searching only for concepts with the type "Laboratory Procedure" or "Laboratory or Test Result".

Direct matching to SNOMED CT using the UMLS Metathesaurus returned 1,991 unique SNOMED CT variables (e.g. basal cell carcinoma [UMLS code C0007117] is directly linked to epithelioma basal cell [SNOCT code 275265005] within databases). 56 variables were manually added by the direct query in the SNOMED CT Browser as no direct UMLS to SNOMED CT connection existed. Further, during the manual review, it was observed that some UMLS concepts which had no direct SNOMED CT equivalent could be applicable to a SNOMED CT variable returned for another concept, so the trial count and additional information was attached from both concepts to the single SNOMED CT variable. When filtered by variables identified in at least 15 trials out of the entire list, a "master list" was generated containing 318 UMLS concepts and 304 SNOMED CT variables (14 concepts had no correlated SNOMED CT variable). The UMLS concepts found in the "master list" were found in 1491 of the 1512 queried trials, i.e., a trial coverage of 98.6%.

Visualization of The Common Eligibility Criteria SNOMED CT Variables and their hierarchical relations

The highly prevalent eligibility criteria concepts in AD trials are listed in Table 2. Since there exist hierarchical relations among these concepts, an online visualization was also generated for these concepts. Each node in the visualization is a common eligibility criteria concept in AD trials followed by its prevalence. For example, "mental disorder" is a node with prevalence of 82.21% because it is used by 82.21% of AD trials for patient screening. The visualization of "master list" concepts and their super classes can be observed at http://htmlpreview.github.io/?https://github.com/Butler925/Alz_viz/blob/master/index_git.htm.

Table 2. The most commonly adopted eligibility criteria variables and their prevalence in AD trials (the last column with column header as "#" indicates the number of parent concepts)

SNOMED-CT Concept Representation for Commonly Adopted Eligibility Variables	SNOMED_ID	Prevalence	Type of	level	Parent_SNOMED ID	#
Clinical finding	404684003	97.09%	finding	1	138875005	1
Disease	64572001	94.25%	disorder	2	404684003	1
Mental disorder	74732009	82.21%	disorder	3	64572001	1
Disorder of brain	81308009	79.50%	disorder	3	64572001	1
Organic mental disorder	111479008	74.74%	disorder	4	74732009, 81308009	2
Dementia	52448006	74.60%	disorder	5	111479008	1
Cerebral degeneration presenting primarily with dementia	279982005	64.62%	disorder	3	64572001	1

Clinical history and observation findings	250171008	64.55%	finding	2	404684003	1
Alzheimer's disease	26929004	64.29%	disorder	6	52448006, 279982005	2
Staging and scales	254291000	60.65%	staging scale	1	138875005	1
Assessment scales	273249006	60.65%	assessment scale	2	254291000	1
Procedure	71388002	58.33%	procedure	1	138875005	1
Observable entity	363787002	51.19%	observable entity	1	138875005	1
Mini-mental state examination	273617000	46.63%	assessment scale	3	273249006	1
Qualifier value	362981000	45.24%	qualifier value	1	138875005	1
General finding of observation of patient	118222006	41.14%	finding	3	250171008	1
Presenile dementia	12348006	39.62%	disorder	6	52448006	1
Disorder of cardiovascular system	49601007	39.55%	disorder	3	64572001	1
Psychological finding	116367006	38.96%	finding	3	250171008	1
Mental state, behavior and/or psychosocial function finding	384821006	38.96%	finding	4	116367006	1
Disorder of nervous system	118940003	35.78%	disorder	3	64572001	1
Current chronological age	424144002	34.06%	observable entity	3	105727008	1
Age AND/OR growth period	105727008	34.06%	observable entity	2	363787002	1
Disorder of blood vessel	27550009	33.33%	disorder	4	49601007	1
Evaluation procedure	386053000	33.33%	procedure	2	71388002	1
Disorder of body system	362965005	32.41%	disorder	3	64572001	1
Cerebrovascular disease	62914000	32.28%	disorder	5	27550009	1
Magnetic resonance imaging	113091000	31.88%	procedure	2	71388002	1
Disorder by body site	123946008	30.16%	disorder	3	64572001	1
Procedure by method	128927009	25.79%	procedure	2	71388002	1
Mood disorder	46206005	25.73%	disorder	4	74732009	1
Substance abuse	66214007	25.66%	disorder	3	64572001	1
Descriptor	272099008	24.80%	qualifier value	2	362981000	1
Cerebrovascular accident	230690007	24.54%	disorder	6	62914000	1
Global assessment of functioning - 1993 Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition adaptation	284061009	23.94%	assessment scale	3	273249006	1
Systemic disease	56019007	23.35%	finding	4	118222006	1
General body state finding	82832008	22.49%	finding	4	118222006	1
Impaired cognition	386806002	21.43%	finding	5	384821006	1
System disorder of the nervous system	230226000	21.16%	disorder	4	118940003	1
Movement disorder	60342002	21.16%	disorder	5	230226000	1
Extrapyramidal disease	76349003	21.16%	disorder	6	60342002	1
Disorder of head	118934005	21.03%	disorder	4	123946008	1
Depressive disorder	35489007	21.03%	disorder	5	46206005	1

SNOMED CT Variable Assessment

Overall, the "master list" contained 21 SNOMED CT semantic types and 13 of the 19 highest-level SNOMED CT variable types. The prevalence of these concepts in AD trials is shown in Table 3, with the top 20 shown in Table 4. Of note, the majority of the variables in Table 4 are specific except for variable "*Disease*", which is very vague. The less vague but still non-specific example variables are "*Systematic Disease*" and "*History of clinical finding in subject*".

Table 3. The counts of trials containing each SNOMED CT semantic type.

SNOMED-CT Semantic Type	Trial Count	Prevalence in Trials

Disorder	1425	94.25%
Finding	1072	70.90%
Assessment scale	917	60.65%
Staging scale	917	60.65%
Procedure	882	58.33%
Observable entity	774	51.19%
Qualifier value	684	45.24%
Situation	250	16.53%
Physical object	231	15.28%
Attribute	163	10.78%
Linkage concept	163	10.78%
Body structure	154	10.19%
Metadata	105	6.94%
Morphologic abnormality	125	8.27%
Mother	56	3.70%
Substance	21	1.39%
Regime/therapy	33	2.18%
Environment	19	1.26%
Environment / location	19	1.26%
Event	17	1.12%
Organism	15	0.99%

Table 4. The top 20 common SNOMED CT terms in AD trials and their prevalence in EHR dataset.

SNOMED CT Term	SNOMED-CT ID	Trial Count	Prevalence in Trials	Count of uses in EHR data for AD patients
Alzheimer's disease	26929004	972	64.29%	30,262
Mini-mental state examination	273617000	705	46.63%	**0**
Presenile dementia	12348006	599	39.62%	7,089
Disease	64572001	555	36.71%	12,029,900
Current chronological age	424144002	515	34.06%	**0**
Mental disorder	74732009	499	33.00%	505,870
Magnetic resonance imaging	113091000	482	31.88%	63,171
Cerebrovascular accident	230690007	371	24.54%	4
Global assessment of functioning - 1993 Diagnostic and Statistical Manual of Mental Disorders- ver.4th	284061009	361	23.88%	**0**
Systemic disease	56019007	353	23.35%	**0**
Disorder of nervous system	118940003	335	22.16%	780,478
Substance abuse	66214007	279	18.45%	9,466
Parkinson's disease	49049000	275	18.19%	**0**
Impaired cognition	386806002	260	17.20%	13,375
Seizure disorder	128613002	240	15.87%	28,586
Hypersensitivity reaction	421961002	218	14.42%	4,686
Schizophrenic disorders	191526005	216	14.29%	40777
History of clinical finding in subject	417662000	207	13.69%	189,543
Risk identification: childbearing family	386414004	205	13.56%	**0**
Clinical dementia rating scale	273367002	204	13.49%	**0**

The Data Gap

Table 5 shows the counts of SNOMED CT variables from the "master list" for each of the five categories. 60% of the variables from the "master list" were found in the EHR dataset, but data for about 40% of the variables that are not available in EHR could be provided by patients without clinicians' assessment. Determining if patients could answer some of the criteria that have no data in the EHR largely relied on health literacy and access to their medical records.

Criteria that are considered symptoms or based on clinical discretion (e.g. amyloid deposition, neurological deficit, psychotic symptom) are unanswerable by patients. Further, specific lab test results (e.g. Cobalamin deficiency, laboratory test abnormal) are also considered to be unanswerable by patients as they may not have the health literacy to address these criteria. Those criteria which are considered answerable by patient are broken into three categories: (1) discrete diagnosis (e.g. Parkinson's Disease, Multiple Sclerosis, Carcinoma of Prostate), (2) answerable with online test (e.g. visual acuity, auditory acuity, memory function), and (3) answerable with structured questions (e.g. Clinical Dementia Rating Scale, Hachinski Ischemia Score, Geriatric Depression Scale). The 'master list' with EHR record counts, descendant record counts, and characterization about how a patient can address the criterion is at https://docs.google.com/spreadsheets/d/1R6_xc_iEq34YUWuJLzT26J1kskEGIGmoQCOgrUJiB3w/edit?usp=sharing.

Table 5. The count of SNOMED CT variables from the "master list" in the five categories.

Category Description	Example	Categories	Total Count
In EHR, categorical variables	*Presenile Dementia*	132	181 (60%)
In EHR, continuous variables	*Laboratory Test*	40	
Not in EHR, can be derived	*Chronological Age*	9	
Not in EHR, answerable by patient	*Questions from Mini-Mental Status Exam*	59	123 (40%)
Not in EHR, not answerable by patient	*General Metabolic Function*	34	
Not applicable	*Psychiatric*	30	

Discussion

The EHR data gap for eligibility screening

From Table 4 we can see that multiple variables used frequently for eligibility screening were not present in the EHR, including *mini-mental state exam questions' answers, global assessment of function, systematic disease, risk identification: child bearing family status,* and *clinical dementia rating scale*. Rating scales used frequently by researchers are usually not available in EHR dataset but constitute important eligibility criteria concepts for AD trials' eligibility criteria. Our study showed that 60.65% of AD trials include assessment scales and 1.79% of AD trials include symptom ratings, whose corresponding data are not available in EHRs.

Overall, forty percent of the "master list" SNOMED CT variables could not be found in the corresponding structured EHR dataset for patients with AD. The percentage is comparable with the 55% coverage of patients' characteristics from the study of Köpcke et al. The two studies' results suggest fully automated EHR-based eligibility screening may still be impossible with the current schema due to the significant data gap, even though both eligibility criteria and EHR data are well represented using a common data model. An improved model may include patient-reported data in areas where criteria are not available in the EHR to allow for comprehensive eligibility criteria coverage.

Patient self-reported data as a new data source

An interesting finding is that 19% of the "master list" SNOMED CT variables did not exist in the EHR but could be answered by patients. The finding suggests the involvement of patients in the eligibility screening process may help recruiting more eligible patients. Successful stories include one by Williams et al.[23] who developed and implemented a computer-assisted interview system in an urban rheumatology clinic, and another by Goncalves et al.[24] who showed that use of patient-facing web forms could capture structured data. However, different opinions also exist. For example, one study by Wuerdeman et al.[25] concluded that patient-reported data are likely not as complete or accurate as the information provided by a provider. Some other barriers also have been reported, such as technological fluency, privacy concerns, and lack of technology infrastructure[26,27]. Further, given that Alzheimer's Disease affects a patient's cognition and often presents in the elderly, this could impact the reliability of patient-reported information so it is important that patient-facing tools would include family members and other stakeholders.

Reusable variables

Since the 304 UMLS concepts from "master list" variables were found in 98.6% of all the Alzheimer's disease clinical trials, the clinical entities associated with these concepts could be adopted as common data elements (CDEs)[28], and may help reducing the workload of future Alzheimer's disease clinical trials by avoiding assessing some of the 9,261 unique clinical entities. There is no currently established CDE for Alzheimer's Disease, so the results of this study could serve as an important first step.

Major Eligibility Criteria

A similar approach to determine the most relevant eligibility criteria was undertaken by using an interview-style approach[11]. Paulson & Weng highlighted the importance of identifying major criteria in creating an optimal clinical trials recruitment tool. Providing equal weight to each eligibility criterion does a disserve in requiring excessive resources for a diminishing return in screening power, so focusing on those most frequent or more important criteria that allow for more robust eligibility screening provides a very strong advantage.

Limitations

This study has multiple limitations. First, only Alzheimer's Disease clinical trials and SNOMED CT variables were included in this study, and this may result in bias in the coverage estimation. If more diseases and all terminologies from OMOP CDM model were included, the assessment of the information gap between EHR and eligibility criteria would be more accurate. Second, we identified a few discrepancies in our SynPUF dataset which may have impacted our results. For example, *Parkinson's Disease* as referenced in the SNOMED CT database found no record counts in patient records, however the dataset used in this analysis identified overlap of Parkinson's Disease in our dataset when searched outside of the SNOMED database. It is possible that there is a coding issue with our dataset, but the more likely scenario is that *Parkinson's Disease* is primarily codified using a different clinical database. Future analyses into data source heterogeneity should also be conducted in an attempt to simplify and centralize how all of this data is referenced. Third, variables such as *Cerebrovascular accident* requires semantic inference and cannot be aligned literally because EHR data may contain specific incidents of Cerebrovascular accident, not this generic concept. Our current simple approach for aligning concepts in criteria and EHR data was unfortunately unable to find its counterpart in the EHR dataset. One implication of this finding is that we need more sophisticated methods for concept matching that is based on semantic alignment between terms, not just based on term matching. Alternative NLP systems to MetaMap, including MedLEE and cTAKES among others, have shown improved identification of clinical terms and may be used in the future to improve on the results elucidated here.[29]

Lastly, one of the most significant limitations in this study involves the intensive manual review necessary to produce these results and its impact on scalability. As evidenced by the 3,610 manual changes made to the original term list in additional to subsequent type modifications and proof-reading, there is a high level of heterogeneity in clinical terminology found in clinical trial eligibility criteria. This heterogeneity increases the workload associated with performing analyses like this and reduces the confidence in the ultimate results. Further, it reduces the scalability of the methods used here. However, tracking of these manual changes does provide some insight into how to address this heterogeneity. Two of the three most common causes for manual modification, formatting and multiple terms, could be easily addressed by using standard term sets or CDEs as mentioned previously. Standardized lists of terms to be used in Alzheimer's Disease eligibility criteria would avoid any variation in terms based on formatting discrepancies and would allow for simple handling of multiple term concepts (e.g. could identify basal cell carcinoma and squamous cell carcinoma is both terms existed in a standard list). Manual modifications due to simplification were performed primarily for the simplicity of this analysis, so future studies into addressing term heterogeneity should also focus on this reason for modification.

Conclusions

We found 40% of the most commonly used criteria variables in Alzheimer's trial are not available in the concept space in EHR of the patients with Alzheimer's disease. The result suggests that EHR-based eligibility screening may not achieve perfect performance due to the information gap. To overcome this challenge, a possible solution could be asking patients for missing information during recruitment when using EHR data for trial-eligible patient screening.

Acknowledgements

This research is supported by grant **2R01LM009886-08A1** (PI: Chunhua Weng). Author AB is also supported by grant 5T35HL007616-37 (PI: Rudolph L. Leibel).

References

1. Gul RB, Ali PA. Clinical trials: the challenge of recruitment and retention of participants. *Journal of Clinical Nursing.* 2010;19(1-2):227-233.
2. Penberthy LT, Dahman BA, Petkov VI, DeShazo JP. Effort Required in Eligibility Screening for Clinical Trials. *Journal of Oncology Practice.* 2012;8(6):365-370.
3. Joseph RR. Viewpoints and concerns of a clinical trial participant. *Cancer.* 1994;74(S9):2692-2693.
4. Campillo-Gimenez B, Buscail C, Zekri O, et al. Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. *Trials.* 2015;16.

5. Joseph G, Dohan D. Recruiting minorities where they receive care: Institutional barriers to cancer clinical trials recruitment in a safety-net hospital. *Contemp Clin Trials.* 2009;30(6):552-559.
6. Penberthy L, Brown R, Puma F, Dahman B. Automated matching software for clinical trials eligibility: Measuring efficiency and flexibility. *Contemporary Clinical Trials.* 2010;31(3):207-217.
7. Ni Y, Wright J, Perentesis J, et al. Increasing the efficiency of trial-patient matching: automated clinical trial eligibility Pre-screening for pediatric oncology patients. *BMC Medical Informatics and Decision Making.* 2015;15.
8. Köpcke F, Lubgan D, Fietkau R, et al. Evaluating predictive modeling algorithms to assess patient eligibility for clinical trials from routine data. *BMC Medical Informatics and Decision Making.* 2013;13(1):134.
9. Thadani SR WC, Bigger JT, Ennever JF, Wajngurt D. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc.* 2009;16(6):869-873.
10. Weng C, Batres C, Borda T, et al. A Real-Time Screening Alert Improves Patient Recruitment Efficiency. *AMIA Annual Symposium Proceedings.* 2011;2011:1489-1498.
11. Paulson ML, Weng C. Desiderata for Major Eligibility Criteria in Breast Cancer Clinical Trials. *AMIA Annu Symp Proc.* 2015;2015:2025-2034.
12. Kang T, Zhang S, Tang Y, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *Journal of the American Medical Informatics Association : JAMIA.* 2017.
13. Si Y, Weng C. An OMOP CDM-based Relational Database of Clinical Research Eligibility Criteria. Medinfo 2017; 2017; Hangzhou, China
14. Overhage JM RP, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc.* 2011;19(1):54-60.
15. Hripcsak G DJ, Shah NH, Reich CG, Huser V, Schuemie MJ, Suchard MA, Park RW, Wong IC, Rijnbeek PR, Van Der Lei J. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform.* 2015;216:574–578.
16. McCarty CA CR, Chute CG, Kullo IJ, Jarvik GP, Larson EB, Li R, Masys DR, Ritchie MD, Roden DM, Struewing JP. The eMERGE network: a consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Med Genomics.* 2011;4(1).
17. Stearns MQ PC, Spackman KA, Wang AY. SNOMED clinical terms: overview of the development process and project status. Proceedings of the AMIA Symposium; 2001.
18. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: A literature review. *Journal of biomedical informatics.* 2010;43(3):451-467.
19. Varghese J, Dugas M. Frequency analysis of medical concepts in clinical trials and their coverage in MeSH and SNOMED-CT. *Methods of information in medicine.* 2015;54(1):83-92.
20. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 2004;32(suppl_1):D267-270.
21. Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings of the AMIA Symposium.* 2001:17-21.
22. Aronson A, Lang F. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc.* 2010;17(3):229-236.
23. Williams CA TT, Mosley-Williams AD. Usability of a computer-assisted interview system for the unaided self-entry of patient data in an urban rheumatology clinic. *J Am Med Inform Assoc.* 2004;11(4):249-259.
24. Gonçalves RS, Tu SW, Nyulas CI, Tierney MJ, Musen MA. An ontology-driven tool for structured data acquisition using Web forms. *Journal of Biomedical Semantics.* 2017;8(1):26.
25. Wuerdeman L VL, Pizziferri L, Tsurikova R, Harris C, Feygin R, Epstein M, Meyers K, Wald JS, Lansky D, Bates DW. How accurate is information that patients contribute to their electronic health record? AMIA Annual Symposium Proceedings 2005; Washington, DC.
26. Archer N, Fevrier-Thomas U, Lokker C, McKibbon KA, Straus SE. Personal health records: a scoping review. *Journal of the American Medical Informatics Association : JAMIA.* 2011;18(4):515-522.
27. Liu LS, Shih PC, Hayes GR. Barriers to the adoption and use of personal health record systems. Proceedings of the 2011 iConference; 2011; Seattle, Washington, USA.
28. Sheehan J, Hirschfeld S, Foster E, et al. Improving the value of clinical research through the use of Common Data Elements. *Clinical Trials.* 2016;13(6):671-676.
29. Wu Y, Denny JC, Rosenbloom ST, Miller RA, Giuse DA, Xu H. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. *AMIA Annual Symposium Proceedings.* 2012;2012:997-1003.