ELaPro, a LOINC-Mapped Core Dataset for Top Laboratory Procedures of Eligibility Screening for Clinical Trials

Ahmed Rafee (ahmed.rafee@outlook.de)
University Hospital of Münster
Sarah Riepenhausen
University of Münster
Philipp Neuhaus
University of Münster
Alexandra Meidt
University of Münster
Martin Dugas
Heidelberg University Hospital
Julian Varghese
University of Münster

Research Article

Keywords: Eligibility Screening, UMLS, LOINC, Data Models, Medical Informatics

Posted Date: December 15th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-1132382/v1

Abstract

Background

Screening for eligible patients continues to pose a great challenge for many clinical trials. This has led to a rapidly growing interest in standardizing computable representations of eligibility criteria (EC) in order to develop tools that leverage data from electronic health record (EHR) systems. Although laboratory procedures (LP) represent a common entity of EC that is readily available and retrievable from EHR systems, there is a lack of interoperable data models for this entity of EC. A public, specialized data model that utilizes international, widely-adopted terminology for LP, e.g. LOINC, is much needed to support automated screening tools.

Objective

The aim of this study is to establish a core dataset for LP most frequently requested to recruit patients for clinical trials using LOINC terminology. Employing such a core dataset could enhance the interface between study feasibility platforms and EHR systems and significantly improve automatic patient recruitment.

Methods

We used a semi-automated approach to analyze 10516 UMLS-annotated screening forms from the Medical Data Models (MDM) portal's data repository. An automated semantic analysis based on concept frequency is followed by a manual expert review performed by physicians to analyze complex recruitment-relevant concepts not amenable to automatic approach.

Results

Based on analysis of 138225 EC from 10516 screening forms, 55 laboratory procedures represented 77.87% of all UMLS laboratory concept occurrences identified in the selected EC forms. We identified 26413 unique UMLS concepts from 118 UMLS semantic types and covered the vast majority of MeSH disease domains.

Conclusions

Only a small set of LP cover the majority of laboratory concepts in screening EC. The results prove the feasibility of establishing a core dataset for a group of LP common to most EC forms. We present ELaPro (Eligibility Laboratory Procedures), a novel, LOINC-mapped, core dataset for the most frequent 55 LP requested in screening for clinical trials in multiple machine-readable data formats.

Introduction

Clinical trials are essential to advance clinical health care and evidence-based medicine [1, 2]. Efficient identification and recruitment of eligible participants is considered a key factor to the success of clinical trials [3–5] and one of its major challenges throughout the last decades [6–9]. Delayed or poor recruitment of target participants in stipulated time remains an enduring problem that leads to increased study costs and reduced power of clinical trials [10–13]. Insufficient participant recruitment is one of the leading causes of early study termination and wasted research resources [14–20].

Eligibility screening is considered the cornerstone of participant recruitment and refers to applying eligibility criteria (EC) to specify the necessary characteristics of study participants who are eligible to participate in a study [21–23]. The wide adoption of Electronic Health Record (EHR) systems in recent years has resulted in large quantities of patient clinical data being available in electronic form, which led to increased interest in establishing and standardizing computable knowledge representations of EC to develop decision support tools for different research aspects e.g. matching eligible patients to clinical trials [24–26]. However, these efforts are challenged by the unstandardized free-text format of EC [27, 28]. In the last decades, different clinical terminologies have been introduced and used to encode medical concepts of EC [29]. These terminologies provided a computable form of EC despite the lack of common standards among different terminologies [22]. One of the most recognized terminology systems is the Unified Medical Language System (UMLS) [30, 31], which is considered a popular option for annotating EC because of its rich metathesaurus and interoperability with other terminologies [22, 32–36]. Over the last years, various methods and techniques have been produced and applied to extract and transform medical concepts from free text into a computable representation using encoding terminologies and annotating tools. This has enhanced the development of automated research tools that utilize patient data from repositories of EHRs to recruit patients for clinical trials [37–48].

Laboratory criteria represent one of the most common categories of EC in clinical trials [49]. There is an obvious lack of dedicated analyses and specialized data models of screening LP. A public, specialized data model in interoperable terminologies for laboratory concepts, e.g. the widely-adopted, international reference of laboratory standards named Logical Observation Identifiers Names and Codes (LOINC®) terminology [50, 51], is much needed to boost computer-based decision support for automated screening for clinical trials. Ross et al. [52] randomly selected 1000 studies from ClinicalTrials.gov and found that laboratory and diagnostic tests represent around 23% of EC in these studies. In 2013 Bhattacharya et al. [53] showed that the semantic type “Diagnostic and Lab Results” constitute the majority of inclusion criteria in both full-text and protocols of ClinicalTrials.gov. Wang et al. [54] classified laboratory and demographic EC to be among the easiest criteria to support automated queries to data repositories from EHRs. Both domains possess a key advantage over other EC domains, in which they are more structured and easy to retrieve from a laboratory information system to support patient recruitment. While many core data models for demographic EC already exist, e.g. Clinical Data Acquisition Standards Harmonization (CDASH) [55], there is a clear research gap when it comes to specialized analyses and data models for LP in eligibility screening. Weng et al. [56] extracted a list of 115 frequent tags from EC

text of 137,889 clinical trials by applying a pure Natural Language Processing (NLP) approach. Doods et al. [57] applied expert-knowledge to manually analyze 17 clinical trials and data elements of EHR systems of several hospitals in Europe to introduce a data inventory of 75 frequent research medical concepts that are available as data items in EHR systems. Kury et al. [58] introduced a dataset of most frequent medical concepts of EC from 1000 random clinical trials divided into 8 semantic domains using manual annotation of concepts into a web-based tool followed by an NLP analysis approach.

Most of earlier work have utilized NLP methods to provide a general approach to semantic domains of EC with very little to no focus on LP. In 2019 Fraser et al. [59] used 3 pre-trained datasets to study the performance of NLP approaches including “Deep Learning” methods in entity recognition, which is essential when studying fine-grained entities of EC like LP. Several methods performed poor (F1 Score = 0.63) on the largest dataset, “MedMentions” [59, 60], that contains over 4000 biomedical abstracts, annotated for UMLS semantic types, suggesting potential challenges when solely applying current NLP techniques to real-world data in the absence of a manual expert review. Recent NLP-based systems like Criteria2Query [61] and ElilE [62] achieved relatively better F1 scores in entity recognition (up to 0.795 and 0.79, respectively). A more laboratory oriented system called Valx [63] showed an F1 score above 0.97, however, this was only tested on a small entity (Diabetes Mellitus I and II).

Clinical terminologies like UMLS have a complex semantic structure with redundant or duplicate concepts, especially when it comes to components of laboratory tests. For instance, the same text string “Albumin” could refer to six semantically different concepts in UMLS (“Biologically Active Substance”, “Amino Acid”, “Peptide, or Protein”, “Gene or Genome”, “Laboratory Procedure”, “Clinical Attribute” or “Physiologic Function”). This issue poses challenges for automatic semantic processing of EC. In addition, many EC use names of pathologic conditions that imply the need to perform a LP rather than directly mentioning the name of the component to be tested, e.g. “Leukocytosis”, which refers to elevated white blood cells in blood belongs to semantic type “Finding”, yet implies the need to perform an LP to fulfill the criteria. Physician-based curations ensure semantic correctness of mapping text strings from EC to clinically relevant laboratory concepts defined in medical terminologies.

This dataset was created by analyzing 138225 EC extracted from 10516 UMLS-annotated screening forms of random clinical trials registered on ClinicalTrials.gov and covering a broad range of different clinical domains (Figure 3). The forms used in this analysis were obtained from the data repository of the Medical Data Models (MDM) portal [64, 65], the largest, non-profit academic portal in Europe for structured medical forms semantically enriched by UMLS annotation performed by medical experts [65]. This semantic enrichment facilitated automated semantic analysis of the included forms and represented a key advantage over analyzing unannotated EC from ClinicalTrials.gov.

In this study, we introduce ELaPro (Eligibility Laboratory Procedures), a novel, public, LOINC-mapped, core dataset of the most frequent LP in screening for clinical trials. We use a semi-automated approach that combines an automated UMLS-based semantic analysis of laboratory concepts followed by a thorough manual expert review. The scope of this analysis is confined to LP following the definition of a

“Laboratory Procedure” by The National Cancer Institute (NCI) metathesaurus [66], which is defined as “A medical procedure that involves testing a sample of blood, urine, or other substance from the body”. Other diagnostic procedures, e.g. radiographic or endoscopic procedures are beyond the scope of this work. ELaPro is an interoperable data model, available in multiple machine-readable formats to be utilized in developing automated screening tools that can be integrated in EHR systems to enhance the recruitment process using real-time queries applied to data repositories of EHR systems.

Methods

Data Collection

A direct access to the local UMLS database (2021AA) as well as the Metadata Repository (MDR) [67], the main database of the MDM portal, was granted by the Institute of Medical Informatics of the University of Muenster for the purpose of this analysis. 10516 eligibility screening forms were identified and included in this analysis. An R-based tool was developed and used to directly access and filter EC forms of MDM portal database and connect them to their UMLS-annotated concepts, which is the “raw data of the automated semantic analysis. tool first filters the ID’s of active forms (latest, most up-to-date annotated version) labeled as “Eligibility Determination” and then specifically filters screening EC forms and extracts all UMLS-annotated concepts from these forms. A list of names and DOI’s of all included EC forms on MDM portal is found in Appendix 1.

Data Analysis

Semantic Form Annotation

Typically, an EC form consists of 2 item groups; Inclusion Criteria and Exclusion Criteria. Each item group consists of items; each item represents a complete element (criterion) of inclusion or exclusion criteria. All medical concepts of each item (criterion) are coded (annotated) using UMLS codes to standardize the representation of free-text EC. The annotating process is performed by a medical expert and reviewed by a physician experienced in UMLS. The detailed process and workflow of the coding process have been thoroughly described in previous works [68-70].

Automated Semantic Analysis in R

The automated part is based on an R-based tool to facilitate extraction and analysis of UMLS codes and their semantic types from pre-annotated screening forms in MDR (n=10516) and the UMLS database. We performed an automated semantic analysis 10516 eligibility screening forms available on the MDM portal as of August 2021. After a thorough study of the structure of the MDR, it was possible to automatically filter EC forms using a unique ID assigned to all forms of eligibility determination in the MDR. Using names and subheadings of forms, we were able to specifically filter the IDs of screening EC forms and eliminate other unwanted types of EC (e.g. follow-up, randomization or continuation criteria). Once the IDs of screening EC forms were collected, an automated retrieval of all UMLS codes used to

annotate medical concepts in these forms was performed. The next step was to automatically measure the frequency of occurrences (n) of these codes and to sort them according to frequency in a descending order.

In order to define the UMLS codes, we used a UMLS table for concept names and sources (MRCONSO) [71]. A subset of MRCONSO that includes the single Preferred Term [71] of each code was created, this aims to identify the single preferred definition of UMLS concepts. Furthermore, the semantic type of each code was automatically identified utilizing the UMLS table for semantic types (MRSTY) [71]. Figure 1 illustrates the process of automated data collection and analysis in both MDM and UMLS databases.

In order to refine results and extract UMLS codes related to laboratory concepts, we needed to define reference semantic types that represent all laboratory concepts in UMLS metathesaurus. Based on prestudy communication with a senior scientist from the National Library of Medicine (NLM) as well as the definition of semantic types, two UMLS semantic Types, “Laboratory Procedure” and “Laboratory or Test Result”, were considered the two reference semantic types for laboratory tests in the UMLS metathesaurus.

Based on these 2 semantic types, results were divided into 2 groups; Group A was assigned the name “EC Laboratory Codes” and includes concepts (codes) from the 2 reference laboratory semantic types mentioned above, while group B was named “Non-Laboratory EC Codes” and includes codes from all other UMLS semantic types. Group B is necessary to ensure that relevant laboratory concepts, which are not linked to the aforementioned semantic types, are still considered for expert review (e.g. concepts like “Leukocytosis” or “Hemoglobin Increased” and many other concepts of semantic type “Finding”). Absolute frequency (n) was automatically counted for all codes in both groups, concepts were then sorted by absolute frequency in a descending pattern from the most frequent (highest n) to the least frequent (lowest n). Figure 1 is a schematic representation of the semi-automated method of this analysis. A list of unique UMLS concepts of group A and B sorted by frequency is found in Appendix 2A and 2B, respectively. A list of all original EC Questions for all codes in group A and B is found in Appendix 3.

Manual Expert Review of Laboratory Concepts

A laborious manual review was necessary to identify and analyze complex concepts that indirectly imply a LP but do not have a laboratory semantic type, thus not amenable to the above mentioned automated semantic analysis. The manual analysis was performed by 2 medical professionals (AR, JV) using Microsoft Excel. If a concept was ambiguous or in doubt it was discussed with 2 additional physicians experienced in UMLS (MD, SR) to decide whether a concept is relevant to a LP or not. We used terms like primary laboratory concept (PLC) and secondary laboratory concept (SLC) to deal with classic issues of UMLS like redundancy (similar, but not identical concepts)) and semantic complexity to help determine the actual representation (nTotal) of laboratory concepts. We provide examples for both in the following two sections.

Primary Laboratory Concept (PLC): refers to the UMLS concept that represents the preferred definition of each laboratory test in the master file. The decision of choosing the UMLS code representing each PLC was made by agreement of 4 physicians. By definition, a PLC must belong to semantic type "Laboratory Procedure" and, if applicable, be as general as possible to accommodate the different standards of the test among different clinical institutes. A PLC for a certain laboratory test is preferably, however not necessarily, the most frequent code among all codes representing that concept. For example, the concept “Creatinine Measurement in Blood” (n = 2) is considered the PLC for creatinine measurement despite having clearly less occurrence frequency than other more specific concepts like “Creatinine Measurement in Serum” (n = 1492) and “Creatinine Measurement in Plasma” (n = 142), since the former is more general and represents other possible variants of the test that might be used in different clinical research institutes.

Secondary Laboratory Concept (SLC): refers to UMLS concepts relevant to a PLC, i.e. it directly or indirectly refers to or implies the same laboratory test component. SLCs include concepts from laboratory semantic types (group A), that are synonymous to a PLC (sibling) as in the previous example of Creatinine, or more typically include concepts from semantic type “Finding”, which usually implies that a test is necessary to evaluate this finding, e.g. “Platelet Count Normal” or “Increased Number of Platelets” imply the need to perform the test, and are therefore secondary to the PLC “Platelet Count Measurement”. SLCs also include certain pathologic conditions that imply the need for a test, e.g. “Hyperkalemia” was considered an SLC to “Blood Potassium Measurement”, “Leukocytosis” is secondary to “White Blood Cell Count Procedure” and “Anemia” is secondary to “Hemoglobin Measurement”, etc. In some rare instances, concepts that referred to a simple relation between two measurable laboratory tests were also considered an SLC if the PLC was part of the ratio, e.g. the concept “Alanine Aminotransferase (ALT) to Aspartate Aminotransferase (AST) Ratio Measurement” was counted with both “ALT Measurement” and with “AST Measurement”.

The Manual Curation (Expert Review): the most common concepts in the laboratory group A (PLCs) were identified based on the frequency of individual occurrence (n), then both A and B groups were searched to find all relevant concepts (SLCs) that directly or indirectly imply the same LP as each of the PLCs. The PLC and its SLCs are then grouped together in a master file to represent one LP (see Figure 2). This process was repeated for each LP identified in group A. Therefore, the results of the manual analysis (master file) include multiple groups of codes, each group represents one LP and is composed of one PLC and multiple SLCs. For each LP, a total count of frequency (nTotal) was calculated by adding all concept occurrences (n) of single codes in the group representing the LP. A “Rank” was assigned to each LP based on its nTotal. The most frequent PLC (highest nTotal) was given rank number 1, second most frequent was given rank number 2 and so on. Unspecific concepts in group A (e.g. “Assay” or “Laboratory Results”) were excluded since they did not refer to any specific test component. Figure 2 shows an example of a manually analyzed LP. A diagram illustrating the manual process is in Appendix 4.

Mapping to LOINC®

The mapping process was based on matching the PLC to a LOINC “COMPONENT”, which is the part of LOINC that specifies what is being measured, evaluated or observed. For most LP, a primary and one or more secondary LOINC codes were assigned. The decision of choosing primary and secondary LOINC concepts was based mainly on a well-recognized core dataset created by the Medical Informatics Initiative (MI-I) [72,74] that includes primary and secondary LOINC concepts for the top 300 most common laboratory tests based on data from 5 different university hospitals in Germany. If no proper matching component was found in the MII core dataset for any of our results, the LOINC database V2.7 was directly used to manually assign a primary, and if applicable a secondary LOINC concept(s). The final step was using the UMLS database to create a core dataset with full LOINC details (component, property, system, etc.) using the LOINC codes mapped to our results and an R-based tool.

Results

Overview

A total of 10516 screening forms containing a total of 138225 criteria were recruited in this study. The MDM portal provides item group names to identify inclusion and exclusion criteria. 20346 item groups within the 10516 forms were identified, 9684 of these item groups (47.59%) were inclusion criteria, while 9727 (47.8%) were exclusion criteria. 932 item groups (4.59%) were unspecifically labeled.

Representation of medical specialties among EC forms (in MeSH® Terms)

Medical Subject Headings (MeSH) [75] is the NLM vocabulary thesaurus used for indexing articles for PubMed [76]. The MDM portal provides a MeSH-based keyword system. Using this system, an automated analysis of representation of broad disease entities and medical specialties among EC forms was performed. We identified 23 unique MeSH subcategories (n = 17340) among included EC forms. “Neoplasms” represent 19.49% (n = 3381) as the most common disease entity among EC forms. Figure 3 shows the distribution of MeSH disease entities among EC forms. Appendix 5 shows absolute frequencies of each MeSH categories.

UMLS Semantic Types in Screening Eligibility Criteria Forms

A total of 27055 unique codes were obtained from the included EC forms, among which 26413 unique UMLS codes (97.62%) were filtered and included in this analysis. These 26413 UMLS codes were used 495516 times and belong to 118 unique semantic types with the most common 5 semantic types being “Finding”, “Disease or Syndrome”, “Pharmacologic Substance”, “Therapeutic or Preventive Procedure” and “Neoplastic Process” based on the frequency of occurrence of codes belonging to these semantic types (n = 3849, 3047, 2201, 2140 and 1204, respectively). Semantic type “Laboratory Procedure” ranked 6th and was one of the top 10 semantic types of UMLS concepts (n = 845), while semantic type “Laboratory or Test Result” was less frequently used (n = 331) and ranked 21st. Concepts from both reference laboratory semantic types combined comprised 4.45% (n = 1176) of all UMLS codes. Appendix 6 shows the frequency of occurrence of UMLS codes of all 118 semantic types.

Laboratory Concepts and Cumulative Frequencies

A total of 58 primary LP (PLCs) were identified by aggregating all relevant (secondary) concepts and calculating nTotal for each LP by adding all frequencies of individual concepts (n) of PLC and its relevant SLC. Rank was assigned to PLCs based on nTotal. The cumulative sum of nTotal was continuously plotted and observed as the concepts were being analyzed (see Figure 4). Analyzed laboratory concepts that have an nTotal above 50 covered the complete transition and the steepest change of the slope of cumulative total frequencies (Orange graph in Figure 4). Based on this, we have only included the first 55 analyzed laboratory concepts (PLCs) that have an nTotal above 50.

The final results of the semi-automated analysis included 55 PLCs as well as 648 SLCs and comprised 703 unique UMLS concepts (2.66% of the 26413 total unique concepts in all included EC forms). Among the 703 unique laboratory concepts included in our final analysis, we identified 311 unique concepts that belong to the group of laboratory semantic types (Group A). These 311 concepts comprised 26.23% of the 1176 concepts in group A and covered 77.87% of its total occurrences (n = 15230/19558). The plot in Figure 5 shows the cumulative frequency of UMLS concepts in the group of laboratory semantic types (Group A) in terms of simple frequency (n). The complete table of results of the manual analysis of PLCs and SLCs can be found in Appendix 7.**Top UMLS Laboratory Concepts in Screening Eligibility Criteria Forms**

The most frequent UMLS laboratory concept in our analysis of screening EC forms was Measurement of Creatinine in Blood with an nTotal of 1817. Table 1 shows a list of the top 55 UMLS laboratory concepts in screening EC forms.

Table 1: Top 55 screening LP ranked according to total frequencies. LP are listed using UMLS and LOINC Terminologies.

Rank	1ry LOINC Code	UMLS Lab. Procedure	UMLS Definition	nTotal
1	59826-8	C3525719	Measurement of creatinine in blood	1817
2	7918-6	C3714540	HIV Antibody Measurement	1670
3	13955-0	C0201487	Hepatitis C antibody measurement	1595
4	76625-3	C0201836	Alanine aminotransferase measurement	1464
5	63557-3	C0201477	Hepatitis B surface antigen measurement	1462
6	2106-3	C0546577	HCG Pregnancy Test	1383
7	26515-7	C0032181	Platelet Count measurement	1381
8	1920-8	C0201899	Aspartate aminotransferase measurement	1303
9	54363-7	C0201913	Bilirubin, total measurement	1265
10	4548-4	C0474680	Hemoglobin A1c measurement	1021
11	59260-0	C0518015	Hemoglobin measurement	968
12	26511-6	C0948762	Absolute neutrophil count	851
13	2164-2	C0373595	Creatinine clearance measurement	758
14	15074-8	C0392201	Blood glucose measurement	633
15	69405-9	C3811844	Estimated Glomerular Filtration Rate	513
16	26464-8	C0023508	White Blood Cell Count procedure	487
17	19197-3	C0201544	Prostate specific antigen measurement	471
18	77145-1	C0428568	Fasting blood glucose measurement	400
19	5010-4	C1533728	Hepatitis C virus genotype determination	352
20	72383-3	C5189164	HER2 in tissue by immunoassay	324
21	1783-0	C0201850	Alkaline phosphatase measurement	287
22	14130-9	C3811131	Estrogen Receptor Measurement	283
23	10676-5	C1868902	HCV viral load	257
24	70218-3	C0202236	Triglycerides measurement	195
25	6298-4	C0729816	Blood potassium measurement	194
26	34714-6	C0525032	International Normalized Ratio	191
27	40557-1	C0373717	Progesterone receptor assay	180
28	1996-8	C0201925	Calcium measurement	171

29	5964-2	C0033707	Prothrombin time assay	137
30	14913-8	C0853134	blood testosterone measurement	131
31	54347-0	C0201838	Albumin measurement	130
32	14647-2	C0201950	Cholesterol measurement test	130
33	22748-8	C0202117	Low density lipoprotein cholesterol measurement	126
34	69739-1	C0202274	Urine drug screen	124
35	26446-5	C2697913	Leukemic Blast Count	122
36	1986-9	C0202100	Insulin C-peptide measurement	118
37	13954-3	C3835873	Serum Hepatitis B E Antigen, qualitative	113
38	3015-5	C0202230	Thyroid stimulating hormone measurement	111
39	50564-4	C0042014	Urinalysis	102
40	33763-4	C1533071	N terminal pro-brain natriuretic peptide level	99
41	82904-4	C2074589	chromosome studies Philadelphia	95
42	20570-8	C0018935	Hematocrit procedure	93
43	3173-2	C0030605	Activated Partial Thromboplastin Time measurement	93
44	83098-4	C0202022	Follicle stimulating hormone measurement	83
45	53962-7	C0201539	Alpha one fetoprotein measurement	80
46	10438-0	C3540684	CD20 Expressing Cell Measurement	76
47	76485-2	C0201657	C-reactive protein measurement	74
48	58410-2	C0009555	Complete Blood Count	71
49	42595-9	C3641250	Hepatitis B DNA Measurement	68
50	14646-4	C0428472	Serum HDL cholesterol measurement	63
51	20564-1	C0523807	Oxygen saturation measurement	60
52	30395-8	C0857490	Granulocyte count	58
53	29760-6	C0201916	Bilirubin, direct measurement	53
54	72903-8	C0005845	Blood urea nitrogen measurement	51
55	1992-7	C0201924	Calcitonin measurement	50

Mapping to LOINC and Generation of Core Dataset

The 55 UMLS LP resulted from this analysis were mapped to LOINC terminology as previously explained in the 'Methods'. Using assigned primary and secondary LOINC concepts, a core dataset was created by completing other LOINC details using LOINC database. The core dataset is available in machine-readable ODM and HL7 FHIR files (see Figure 6) in UMLS and LOINC terminologies at https://www.doi.org/10.21961/mdm:44732. CSV, ODM and FHIR formats of the dataset are found in Appendix 8A-8C.

Discussion

Principal Findings

The purpose of this study is to identify the LP most frequently needed to recruit patients for clinical trials and test the feasibility of establishing a core dataset that can be used by tools of automated screening to help improve patient enrollment in clinical trials. The results show that only a small number of LP is frequently requested in most screening EC, clearly more than other LP, an observation that can be clearly seen in the coverage graph depicted in Figure 4 and 5 where 311 UMLS concepts representing only 55 LP covered 77.87% of all laboratory concept occurrences in screening EC forms. These findings clearly confirm the feasibility of creating a core dataset.

The results of this analysis include beside the dataset for LP, another dataset for the complete set of UMLS concepts and their 118 semantic types identified in 10,000+ EC forms. These results could further contribute to the improvement of clinical research by serving as a rich source of data for researchers studying complexity and semantic content of EC. They can also be utilized as raw data to perform further analyses on other semantic domains, which might produce new core datasets that could contribute further to the enhancement of automated screening for clinical trials.

Comparison To Related Data Models

We compare ELaPro to the afore mentioned datasets of Weng et al. [56], Doods et al. [57] and Kury et al. [58]. Weng created a general list of the 115 most common tags in EC without any focus on a certain domain, his list included only 20 concepts that may directly or indirectly refer to a LP, ELaPro overlapped with 17 of these concepts (10 PLCs and 7 SLCs). The data inventory of Doods et al. [57] comprised 41 laboratory concepts in the domain "Laboratory Finding", 29 of which overlapped with the dataset ELaPro. Kury et al. [58] sorted the results into domains, e.g. Device, Condition and Measurement etc. and analyzed the 15 most common concepts of each domain. Laboratory concepts were part of the domain "Measurement" and comprised 9 of its 15 most common tokens. All nine laboratory concepts from Kury's dataset were part of ELaPro. All these findings provide supporting evidences of the accuracy and generalizability of our results.

While most of these relevant studies presented a general analysis of semantic domains of EC, our study introduces a specialized analysis for one entity of EC that is considered common [52–54] and optimal for automated queries of EHR systems [54], i.e., LP, for which a clear gap in research and data models exist. ELaPro is the result of analyzing a large number of UMLS-annotated screening EC forms (19516), thereby clearly exceeding the sample size of many relevant earlier studies. Furthermore, the EC forms used in this analysis covered almost all MeSH disease domains (see Figure 3), which produces more representative results and eliminates the bias that might come from being restricted to a specific clinical domain.

Recent NLP-based systems like Criteria2Query [61], ElilE [62] or Valx [63] provide a more scalable informatics approach. However, our approach puts emphasis on highest accuracy of the results through physician-based curation.

Strengths

Many earlier studies that analyzed EC were based completely or to a large extent on automated approaches like NLP, whose performance in analyzing fine-grained entities might be suboptimal compared to studies that involve manual expert review to process their data models similar to our study. ELaPro is a novel core dataset, not only because it combines an automated approach of semantic analysis followed by an intensive manual expert review to analyze complex EC patterns and concepts, but also because it represents the first public specialized dataset of LP in eligibility screening combining the UMLS with LOINC, one of most widely-adopted international references for laboratory concepts. ELaPro can serve as a data model to enhance the interface between study feasibility platforms and EHR systems to develop automated tools for screening and optimize patient recruitment for clinical trials. This dataset is available in interoperable machine-readable formats like Operational Data Model (ODM) [77] and Fast Health Interoperability Resources (FHIR) [78].

Limitations and Challenges

Approach and Scalability: We realize that our approach does not provide an ultimate solution to automating patient recruitment, mainly due to lacking threshold values and comparable operators. However, providing an expert-curated dataset of the most common laboratory concepts can contribute to automate eligibility screening in many different systems.

The automated part of the analysis facilitates extraction and analysis of UMLS codes from a large number of pre-annotated forms from the MDR of MDM portal. Furthermore, PLC and SLC terms illustrate the importance of manual curation in dealing with the issues of UMLS like redundancy and semantic complexity. While our expert-based approach is not scalable in general, it ensures high accuracy in finding relevant lab concepts in unambiguous text strings, which commonly lack entries and links in the LOINC or UMLS terminology system.

Potential Biases: The vast majority of EC analyzed in this study were originally taken from ClinicalTrials.gov, which could pose a potential bias towards trials from the united states of America.

Furthermore, the distribution of MeSH disease domains among included forms (Figure 3) could also pose a potential bias towards the domains that are relatively more represented, i.e. neoplasms, cardiovascular and immune system diseases. Our work aims to study a representative sample of all major MeSH disease entities to produce a general core dataset of common LP. This set can be applied in EHR systems to enhance participant recruitment for clinical studies in many different domains

The complexity and redundancy of UMLS metathesaurus: ambiguous or duplicate concepts represent a known issue in terminologies like UMLS [79], where one concept can belong to multiple semantic types at the same time, e.g. the concept “Albumin” can be found in both semantic types “Amino Acid, Peptide or Protein” and “Laboratory Procedure”. This has led to many LP being confusingly annotated by redundant concepts with non-laboratory semantic types.This issue was dealt with in the manual part by using the hierarchy of PLC and SLC, where SLC’s represented all the possible redundant concepts that refer to the main laboratory concept (PLC).

Certain EC concepts like “Leukocytosis”, which do not belong to the two main laboratory semantic types, yet indirectly imply the need for a LP, posed a challenge to this study as they are not amenable to automated semantic analysis. In most instances, this problem is solved by the manual expert review performed by a physician, but in some instances, the EC was vague or ill-defined, in which a LP is implied without specifying the exact component to be tested, e.g. “Abnormal Liver Function”, in this case, these concepts were excluded from the analysis.

Conclusion

In this study we present ELaPro, the first specialized public core dataset for the most frequent 55 laboratory procedures in EC to recruit patients for clinical trials. This study proves the feasibility of establishing such a dataset and highlights the importance of using a semi-automated approach. This dataset is mapped to LOINC, one of the most widely-adopted international references of laboratory concepts. ELaPro is available in machine-readable formats like CSV, ODM and HL7 FHIR

And can serve as a blueprint to enhance the interface between study feasibility platforms and EHR systems to develop automated tools of patient recruitment.

Abbreviations

ELaPro	Eligibility Laboratory Procedures
EC	Eligibility Criteria
EHR	Electronic Health Record
LP	Laboratory Procedure(s)
UMLS	Unified Medical Language System
MDM	Medical Data Models
LOINC	Logical Observation Identifiers Names and Codes
n	Frequency of Occurrence
STR	String
SAB	Abbreviated Source Name
CUI	Concept Unique Identifier
STY	Semantic Type
NLM	National Library of Medicine
nTotal	Total count of frequencies
MeSH	Medical Subject Headings
PLC	Primary Laboratory Concept
SLC	Secondary Laboratory Concept
ODM	Operational Data Model
FHIR	Fast Health Interoperability Resources
CSV	Comma-Separated Values

Declarations

Ethics

This study does not contain any studies with animal subjects, human participants or human tissues performed by any of the authors.

Consent to publish

This study does not involve details, images, or videos relating to an individual person. Therefore, no consent is applicable or necessary.

Author Contribution

JV and MD conceived the idea for the study and supervised the workflow. AR developed the automated analysis tool in R, performed the automated and manual analyses, wrote and revised the manuscript and prepared all graphs and tables. SR performed a sample audit for results of automated analysis. PN and AM provided technical and administrative support. All authors read and commented the manuscript and approved the final version.

Funding

All authors confirm that there is no funding to declare in this study.

Competing interest

All authors confirm that they have no conflict of interest associated with this publication.

Availability of Data and material

The eligibility screening forms analyzed in this study are available in the Meta-Data Repository of the Medical-Data-Models Portal https://medical-data-models.org (Category: Eligibility Determination). Generated data are available as downloadable, machine-readable ODM and FHIR files at https://www.doi.org/10.21961/mdm:44732.

References

1. Pung, J., & Rienhoff, O. (2020). Key components and IT assistance of participant management in clinical research: a scoping review. JAMIA open, 3(3), 449–458. https://doi.org/10.1093/jamiaopen/ooaa041
2. Vose, J. M., Chuk, M. K., & Giles, F. (2017). Challenges in Opening and Enrolling Patients in Clinical Trials. American Society of Clinical Oncology educational book. American Society of Clinical Oncology. Annual Meeting, 37, 139–143. https://doi.org/10.1200/EDBK_179807
3. Bower, P., Brueton, V., Gamble, C., Treweek, S., Smith, C. T., Young, B., & Williamson, P. (2014). Interventions to improve recruitment and retention in clinical trials: a survey and workshop to assess current practice and future priorities. Trials, 15, 399. https://doi.org/10.1186/1745-6215-15-399
4. Gardner, H. R., Albarquoni, L., El Feky, A., Gillies, K., & Treweek, S. (2020). A systematic review of non-randomised evaluations of strategies to improve participant recruitment to randomised controlled trials. F1000Research, 9, 86. https://doi.org/10.12688/f1000research.22182.1
5. Zahren, C., Harvey, S., Weekes, L., Bradshaw, C., Butala, R., Andrews, J., & O'Callaghan, S. (2021). Clinical trials site recruitment optimisation: Guidance from Clinical Trials: Impact and Quality. Clinical trials (London, England), 18(5), 594–605. https://doi.org/10.1177/17407745211015924
6. Haidich, A. B., & Ioannidis, J. P. (2001). Patterns of patient enrollment in randomized controlled trials. Journal of clinical epidemiology, 54(9), 877–883. https://doi.org/10.1016/s0895-4356(01)00353-5

7. McDonald, A. M., Knight, R. C., Campbell, M. K., Entwistle, V. A., Grant, A. M., Cook, J. A., Elbourne, D. R., Francis, D., Garcia, J., Roberts, I., & Snowdon, C. (2006). What influences recruitment to randomised controlled trials? A review of trials funded by two UK funding agencies. Trials, 7, 9. https://doi.org/10.1186/1745-6215-7-9
8. Gul, R. B., & Ali, P. A. (2010). Clinical trials: the challenge of recruitment and retention of participants. Journal of clinical nursing, 19(1-2), 227–233. https://doi.org/10.1111/j.1365-2702.2009.03041.x
9. Cullati, S., Courvoisier, D. S., Gayet-Ageron, A., Haller, G., Irion, O., Agoritsas, T., Rudaz, S., & Perneger, T. V. (2016). Patient enrollment and logistical problems top the list of difficulties in clinical research: a cross-sectional survey. BMC medical research methodology, 16, 50. https://doi.org/10.1186/s12874-016-0151-1
10. Team, V., Bugeja, L., & Weller, C. D. (2018). Barriers and facilitators to participant recruitment to randomised controlled trials: A qualitative perspective. International wound journal, 15(6), 929–942. https://doi.org/10.1111/iwj.12950
11. Gardner, H. R., Albarquoni, L., El Feky, A., Gillies, K., & Treweek, S. (2020). A systematic review of non-randomised evaluations of strategies to improve participant recruitment to randomised controlled trials. F1000Research, 9, 86. https://doi.org/10.12688/f1000research.22182.1
12. Chaudhari, N., Ravi, R., Gogtay, N. J., & Thatte, U. M. (2020). Recruitment and retention of the participants in clinical trials: Challenges and solutions. Perspectives in clinical research, 11(2), 64–69. https://doi.org/10.4103/picr.PICR_206_19
13. Houghton, C., Dowling, M., Meskell, P., Hunter, A., Gardner, H., Conway, A., Treweek, S., Sutcliffe, K., Noyes, J., Devane, D., Nicholas, J. R., & Biesty, L. M. (2020). Factors that impact on recruitment to randomised trials in health care: a qualitative evidence synthesis. The Cochrane database of systematic reviews, 10(10), MR000045. https://doi.org/10.1002/14651858.MR000045.pub2
14. Kasenda, B., von Elm, E., You, J., Blümle, A., Tomonaga, Y., Saccilotto, R., Amstutz, A., Bengough, T., Meerpohl, J. J., Stegert, M., Tikkinen, K. A., Neumann, I., Carrasco-Labra, A., Faulhaber, M., Mulla, S. M., Mertz, D., Akl, E. A., Bassler, D., Busse, J. W., Ferreira-González, I., … Briel, M. (2014). Prevalence, characteristics, and publication of discontinued randomized trials. JAMA, 311(10), 1045–1051. https://doi.org/10.1001/jama.2014.1361
15. Briel, M., Olu, K. K., von Elm, E., Kasenda, B., Alturki, R., Agarwal, A., Bhatnagar, N., & Schandelmaier, S. (2016). A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. Journal of clinical epidemiology, 80, 8–15. https://doi.org/10.1016/j.jclinepi.2016.07.016
16. Walters, S. J., Bonacho Dos Anjos Henriques-Cadby, I., Bortolami, O., Flight, L., Hind, D., Jacques, R. M., Knox, C., Nadin, B., Rothwell, J., Surtees, M., & Julious, S. A. (2017). Recruitment and retention of participants in randomised controlled trials: a review of trials funded and published by the United Kingdom Health Technology Assessment Programme. BMJ open, 7(3), e015276. https://doi.org/10.1136/bmjopen-2016-015276

17. Peckham, E., Arundel, C., Bailey, D., Callen, T., Cusack, C., Crosland, S., Foster, P., Herlihy, H., Hope, J., Ker, S., McCloud, T., Romain-Hooper, C. B., Stribling, A., Phiri, P., Tait, E., Gilbody, S., & SCIMITAR+ collaborative (2018). Successful recruitment to trials: findings from the SCIMITAR+ Trial. Trials, 19(1), 53. https://doi.org/10.1186/s13063-018-2460-7
18. Daykin, A., Clement, C., Gamble, C., Kearney, A., Blazeby, J., Clarke, M., Lane, J. A., & Shaw, A. (2018). 'Recruitment, recruitment, recruitment' - the need for more focus on retention: a qualitative study of five trials. Trials, 19(1), 76. https://doi.org/10.1186/s13063-018-2467-0
19. Briel, M., Speich, B., von Elm, E., & Gloy, V. (2019). Comparison of randomized controlled trials discontinued or revised for poor recruitment and completed trials with the same research question: a matched qualitative study. Trials, 20(1), 800. https://doi.org/10.1186/s13063-019-3957-4
20. Fogel D. B. (2018). Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: A review. Contemporary clinical trials communications, 11, 156–164. https://doi.org/10.1016/j.conctc.2018.08.001
21. Van Spall, H. G., Toren, A., Kiss, A., & Fowler, R. A. (2007). Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA, 297(11), 1233–1240. https://doi.org/10.1001/jama.297.11.1233
22. Weng, C., Tu, S. W., Sim, I., & Richesson, R. (2010). Formal representation of eligibility criteria: a literature review. Journal of biomedical informatics, 43(3), 451–467. https://doi.org/10.1016/j.jbi.2009.12.004
23. Kim, E. S., Bernstein, D., Hilsenbeck, S. G., Chung, C. H., Dicker, A. P., Ersek, J. L., Stein, S., Khuri, F. R., Burgess, E., Hunt, K., Ivy, P., Bruinooge, S. S., Meropol, N., & Schilsky, R. L. (2015). Modernizing Eligibility Criteria for Molecularly Driven Trials. Journal of clinical oncology : official journal of the American Society of Clinical Oncology, 33(25), 2815–2820. https://doi.org/10.1200/JCO.2015.62.1854
24. Dugas M, Lange M, Müller-Tidow C, Kirchhof P, Prokosch HU. Routine data from hospital information systems can support patient recruitment for clinical studies. Clin Trials. 2010 Apr;7(2):183-9. doi: 10.1177/1740774510363013. Epub 2010 Mar 25. PMID: 20338903.
25. Weng C. Optimizing Clinical Research Participant Selection with Informatics. Trends Pharmacol Sci. 2015;36(11):706-709. doi: 10.1016/j.tips.2015.08.007
26. O'Brien, E. C., Raman, S. R., Ellis, A., Hammill, B. G., Berdan, L. G., Rorick, T., Janmohamed, S., Lampron, Z., Hernandez, A. F., & Curtis, L. H. (2021). The use of electronic health records for recruitment in clinical trials: a mixed methods analysis of the Harmony Outcomes Electronic Health Record Ancillary Study. Trials, 22(1), 465. https://doi.org/10.1186/s13063-021-05397-0
27. Tu, S. W., Peleg, M., Carini, S., Bobak, M., Ross, J., Rubin, D., & Sim, I. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. Journal of biomedical informatics, 44(2), 239–250. https://doi.org/10.1016/j.jbi.2010.09.007
28. Pressler, T. R., Yen, P. Y., Ding, J., Liu, J., Embi, P. J., & Payne, P. R. (2012). Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-

screening tools. BMC medical informatics and decision making, 12, 47. https://doi.org/10.1186/1472-6947-12-47

29. Dalianis H. (2018). Medical Classifications and Terminologies. Clinical Text Mining: Secondary Use of Electronic Patient Records, 35–43. https://doi.org/10.1007/978-3-319-78503-5_5
30. Unified Medical Language System, Official Website: https://www.nlm.nih.gov/research/umls.
31. Lindberg, D. A., Humphreys, B. L., & McCray, A. T. (1993). The Unified Medical Language System. Methods of information in medicine, 32(4), 281–291. https://doi.org/10.1055/s-0038-1634945
32. Friedman C. (1997). Towards a comprehensive medical language processing system: methods and issues. Proceedings : a conference of the American Medical Informatics Association. AMIA Fall Symposium, 595–599.
33. Patel, C. O., & Cimino, J. J. (2008). Using semantic and structural properties of the UMLS to discover potential terminological relationships. AMIA ... Annual Symposium proceedings. AMIA Symposium, 2008, 555.
34. Patel, C. O., & Weng, C. (2008). ECRL: an eligibility criteria representation language based on the UMLS Semantic Network. AMIA ... Annual Symposium proceedings. AMIA Symposium, 1084.
35. Reimer, A. P., & Milinovich, A. (2020). Using UMLS for electronic health data standardization and database design. Journal of the American Medical Informatics Association : JAMIA, 27(10), 1520–1528. https://doi.org/10.1093/jamia/ocaa176
36. Rasmy, L., Tiryaki, F., Zhou, Y., Xiang, Y., Tao, C., Xu, H., & Zhi, D. (2020). Representation of EHR data for predictive modeling: a comparison between UMLS and other terminologies. Journal of the American Medical Informatics Association : JAMIA, 27(10), 1593–1599. https://doi.org/10.1093/jamia/ocaa180
37. Thadani, S. R., Weng, C., Bigger, J. T., Ennever, J. F., & Wajngurt, D. (2009). Electronic screening improves efficiency in clinical trial recruitment. Journal of the American Medical Informatics Association : JAMIA, 16(6), 869–873. https://doi.org/10.1197/jamia.M3119
38. Penberthy, L., Brown, R., Puma, F., & Dahman, B. (2010). Automated matching software for clinical trials eligibility: measuring efficiency and flexibility. Contemporary clinical trials, 31(3), 207–217. https://doi.org/10.1016/j.cct.2010.03.005
39. Köpcke, F., & Prokosch, H. U. (2014). Employing computers for the recruitment into clinical trials: a comprehensive systematic review. Journal of medical Internet research, 16(7), e161. https://doi.org/10.2196/jmir.3446
40. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, Li Q, Zhai H, Solti I. Automated CT eligibility prescreening: increasing the efficiency of patient identification for CTs in the emergency department. J Am Med Inform Assoc. 2015 Jan;22(1):166-78. doi: 10.1136/amiajnl-2014-002887. Epub 2014 Jul 16. PMID: 25030032; PMCID: PMC4433376.
41. Ni Y, Wright J, Perentesis J, et al. Increasing the efficiency of trial-patient matching: automated CT eligibility pre-screening for pediatric oncology patients. BMC Med Inform Decis Mak. 2015;15:28. Published 2015 Apr 14. doi:10.1186/s12911-015-0149-3.

42. Zhang K, Demner-Fushman D. Automated classification of eligibility criteria in CTs to facilitate patient-trial matching for specific patient populations. J Am Med Inform Assoc. 2017;24(4):781-787. doi:10.1093/jamia/ocw176.
43. Wilson, C., Rooshenas, L., Paramasivan, S., Elliott, D., Jepson, M., Strong, S., Birtle, A., Beard, D. J., Halliday, A., Hamdy, F. C., Lewis, R., Metcalfe, C., Rogers, C. A., Stein, R. C., Blazeby, J. M., & Donovan, J. L. (2018). Development of a framework to improve the process of recruitment to randomised controlled trials (RCTs): the SEAR (Screened, Eligible, Approached, Randomised) framework. Trials, 19(1), 50. https://doi.org/10.1186/s13063-017-2413-6
44. Devoe, C., Gabbidon, H., Schussler, N., Cortese, L., Caplan, E., Gorman, C., Jethwani, K., Kvedar, J., & Agboola, S. (2019). Use of Electronic Health Records to Develop and Implement a Silent Best Practice Alert Notification System for Patient Recruitment in Clinical Research: Quality
45. Gligorijevic, J., Gligorijevic, D., Pavlovski, M., Milkovits, E., Glass, L., Grier, K., Vankireddy, P., & Obradovic, Z. (2019). Optimizing clinical trials recruitment via deep learning. Journal of the American Medical Informatics Association : JAMIA, 26(11), 1195–1202. https://doi.org/10.1093/jamia/ocz064 Improvement Initiative. JMIR medical informatics, 7(2), e10020. https://doi.org/10.2196/10020
46. Meystre, S. M., Heider, P. M., Kim, Y., Aruch, D. B., & Britten, C. D. (2019). Automatic trial eligibility surveillance based on unstructured clinical data. International journal of medical informatics, 129, 13–19. https://doi.org/10.1016/j.ijmedinf.2019.05.018
47. Blatch-Jones, A., Nuttall, J., Bull, A., Worswick, L., Mullee, M., Peveler, R., Falk, S., Tape, N., Hinks, J., Lane, A. J., Wyatt, J. C., & Griffiths, G. (2020). Using digital tools in the recruitment and retention in randomised controlled trials: survey of UK Clinical Trial Units and a qualitative study. Trials, 21(1), 304. https://doi.org/10.1186/s13063-020-04234-0
48. Cai, T., Cai, F., Dahal, K. P., Cremone, G., Lam, E., Golnik, C., Seyok, T., Hong, C., Cai, T., & Liao, K. P. (2021). Improving the Efficiency of Clinical Trial Recruitment Using an Ensemble Machine Learning to Assist With Eligibility Screening. ACR open rheumatology, 3(9), 593–600. https://doi.org/10.1002/acr2.11289
49. Spira AI, Stewart MD, Jones S, Chang E, Fielding A, Richie N, Wood LS, Thompson MA, Jones L, Nair A, Mahal BA, Gerber DE. Modernizing CT Eligibility Criteria: Recommendations of the ASCO-Friends of Cancer Research Laboratory Reference Ranges and Testing Intervals Work Group. Clin Cancer Res. 2021 May 1;27(9):2416-2423. doi: 10.1158/1078-0432.CCR-20-3853. Epub 2021 Feb 9. PMID: 33563636; PMCID: PMC8102342.
50. Huff, S. M., Rocha, R. A., McDonald, C. J., De Moor, G. J., Fiers, T., Bidgood, W. D., Jr, Forrey, A. W., Francis, W. G., Tracy, W. R., Leavelle, D., Stalling, F., Griffin, B., Maloney, P., Leland, D., Charles, L., Hutchins, K., & Baenziger, J. (1998). Development of the Logical Observation Identifier Names and Codes (LOINC) vocabulary. Journal of the American Medical Informatics Association : JAMIA, 5(3), 276–292.

51. Bodenreider, O., Cornet, R., & Vreeman, D. J. (2018). Recent Developments in Clinical Terminologies - SNOMED CT, LOINC, and RxNorm. Yearbook of medical informatics, 27(1), 129–139. https://doi.org/10.1055/s-0038-1667077 https://doi.org/10.1136/jamia.1998.0050276
52. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in CTs. Summit Transl Bioinform. 2010 Mar 1;2010:46-50. PMID: 21347148; PMCID: PMC3041539.
53. Bhattacharya S, Cantor MN. Analysis of eligibility criteria representation in industry-standard CT protocols. J Biomed Inform. 2013 Oct;46(5):805-13. doi: 10.1016/j.jbi.2013.06.001. Epub 2013 Jun 12. PMID: 23770150.
54. AY, Lancaster WJ, Wyatt MC, Rasmussen LV, Fort DG, Cimino JJ. Classifying CT Eligibility Criteria to Facilitate Phased Cohort Identification Using Clinical Data Repositories. AMIA Annu Symp Proc. 2018;2017:1754-1763. Published 2018 Apr 16.
55. Official page of CDASH on CDISC Website: https://www.cdisc.org/standards/foundational/cdash
56. Miotto R, Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. J Biomed Inform. 2013;46(6):1145-1151. doi:10.1016/j.jbi.2013.08.012.
57. Doods J, Botteri F, Dugas M, Fritz F; EHR4CR WP7. A European inventory of common electronic health record data elements for CT feasibility. Trials. 2014 Jan 10;15:18. doi: 10.1186/1745-6215-15-18. PMID: 24410735; PMCID: PMC3895709.
58. Kury F, Butler A, Yuan C, Fu LH, Sun Y, Liu H, Sim I, Carini S, Weng C. Chia, a large annotated corpus of CT eligibility criteria. Sci Data. 2020 Aug 27;7(1):281. doi: 10.1038/s41597-020-00620-0. PMID: 32855408; PMCID: PMC7452886.
59. *Fraser, K.C., Nejadgholi, I., Bruijn, B.D., Li, M., LaPlante, A., & Abidine, K.Z. (2019). Extracting UMLS Concepts from Medical Text Using General and Domain-Specific Deep Learning Models. ArXiv, abs/1910.01274.* https://arxiv.org/abs/1910.01274
60. *Mohan, S., & Li, D. (2019). MedMentions: A Large Biomedical Corpus Annotated with UMLS Concepts. ArXiv, abs/1902.09476.*
61. Yuan, C., Ryan, P. B., Ta, C., Guo, Y., Li, Z., Hardin, J., Makadia, R., Jin, P., Shang, N., Kang, T., & Weng, C. (2019). Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association : JAMIA, 26(4), 294–305. https://doi.org/10.1093/jamia/ocy178
62. Kang, T., Zhang, S., Tang, Y., Hruby, G. W., Rusanov, A., Elhadad, N., & Weng, C. (2017). EliIE: An open-source information extraction system for clinical trial eligibility criteria. Journal of the American Medical Informatics Association : JAMIA, 24(6), 1062–1071. https://doi.org/10.1093/jamia/ocx019
63. Hao, T., Liu, H., & Weng, C. (2016). Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text. Methods of information in medicine, 55(3), 266–275. https://doi.org/10.3414/ME15-01-0112
64. Medical Data Models Portal, Official Website: https://medical-data-models.org.
65. Dugas M, Neuhaus P, Meidt A, Doods J, Storck M, Bruland P, et al. Portal of medical data models: information infrastructure for medical research and healthcare. Database (Oxford)

2016;2016:pii:bav121.

66. National Cancer Institute Metathesaurus, Official Website:https://ncimetathesaurus.nci.nih.gov
67. Hegselmann, S., Storck, M., Gessner, S. et al. Pragmatic MDR: a metadata repository with bottom-up standardization of medical metadata through reuse. BMC Med Inform Decis Mak **21,** 160 (2021). https://doi.org/10.1186/s12911-021-01524-8
68. Varghese J, Dugas M. Frequency analysis of medical concepts in CTs and their coverage in MeSH and SNOMED-CT. Methods Inf Med. 2015;54(1):83-92. doi: 10.3414/ME14-01-0046. Epub 2014 Oct 27. PMID: 25346408.
69. Holz C, Kessler T, Dugas M, Varghese J. Core Data Elements in Acute Myeloid Leukemia: A Unified Medical Language System-Based Semantic Analysis and Experts' Review. JMIR Med Inform. 2019;7(3):e13554. Published 2019 Aug 12. doi:10.2196/13554.
70. Kentgen M, Varghese J, Samol A, Waltenberger J, Dugas M. Common Data Elements for Acute Coronary Syndrome: Analysis Based on the Unified Medical Language System. JMIR Med Inform. 2019;7(3):e14107. Published 2019 Aug 23. doi:10.2196/14107.
71. UMLS® Reference Manual [Internet]. Bethesda (MD): National Library of Medicine (US); 2009 Sep-. 2, Metathesaurus. [Updated 2021 Aug 20]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK9684
72. German Medical Informatics Initiative (MI-I), Official Website: https://www.medizininformatik-initiative.de/en/start
73. Core Dataset of the German Medical Informatics Initiative, Official Webpage:https://www.medizininformatik-initiative.de/en/basic-modules-mii-core-data-set
74. Semler, S. (2019). LOINC: Origin, development of and perspectives for medical research and biobanking – 20 years on the way to implementation in Germany. Journal of Laboratory Medicine, 43(6), 359-382. https://doi.org/10.1515/labmed-2019-0193
75. Medical Subject Headings, Official Webpage: https://www.nlm.nih.gov/mesh/meshhome.html.
76. Official Website of PubMed: https://pubmed.ncbi.nlm.nih.gov.
77. Operational Data Model. [Online] [cited 2014]. Available from: http://www.cdisc.org/odm.
78. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R., & Stiawan, D. (2021). The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. JMIR medical informatics, 9(7), e21929. https://doi.org/10.2196/21929
79. Cimino J. J. (2001). Battling Scylla and Charybdis: the search for redundancy and ambiguity in the 2001 UMLS metathesaurus. Proceedings. AMIA Symposium, 120–124.

Figures

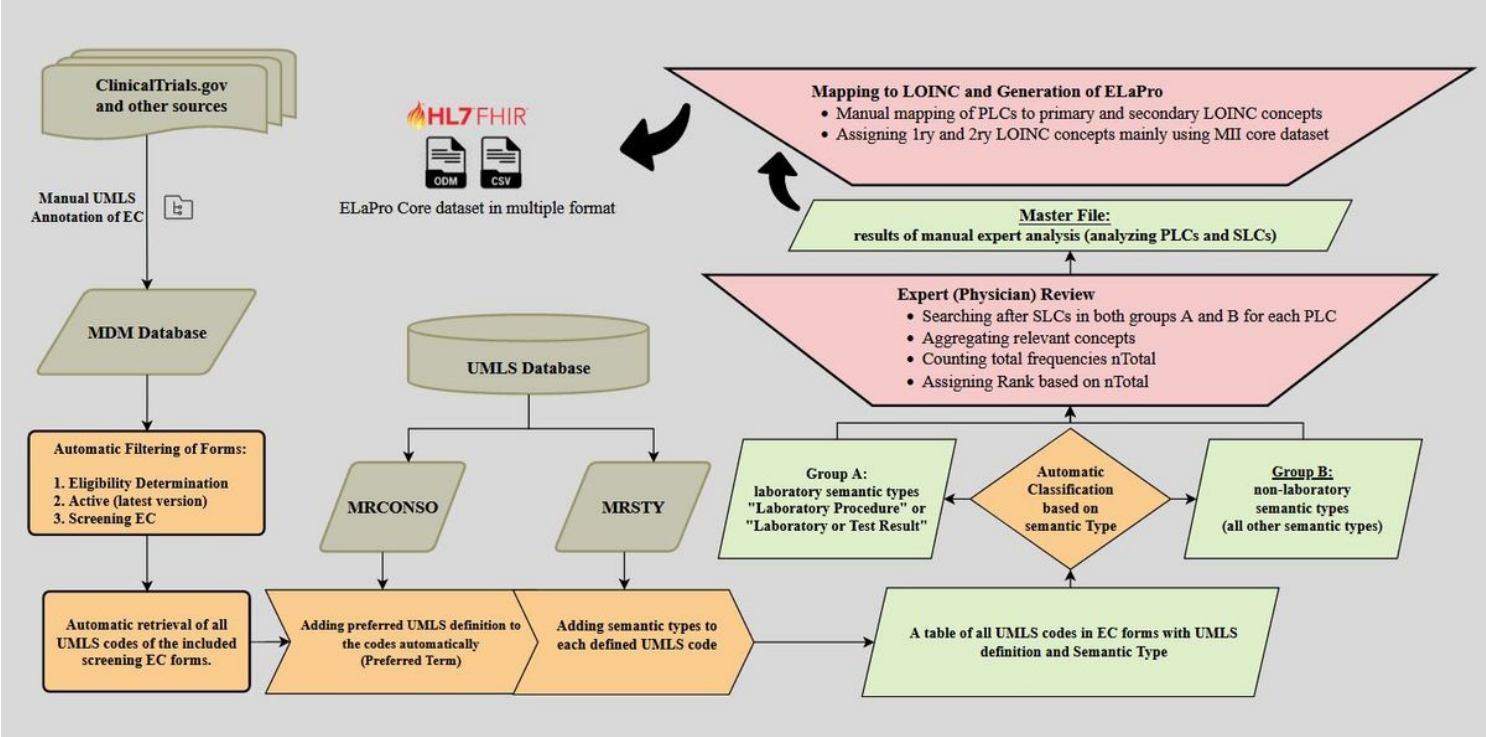

Figure 1

Schematic representation of the semi-automated method used in this analysis. MDR: Metadata Repository; MDM: Medical Data Models Portal; EC: Eligibility Criteria; UMLS: Unified Medical Language System; MRCONSO: a UMLS table for concept names and sources; MRSTY: a UMLS table for semantic types; SLC: Secondary Laboratory Concept; PLC: Primary Laboratory Concept; LOINC: Logical Observation Identifiers Names and Codes; MII: German Medical Informatics Initiative [72,74]; ODM: Operational Data Model; CSV: Comma-Separated Values.

Figure 2

Manually analyzed concept of Bilirubin, Total Measurement. The row at the bottom shows mapping of PLC to LOINC (Further LOINC details were omitted from the image but can still be found in the dataset). PLC: Primary Laboratory Concept; STR: string (definition); n: frequency of individual concept; STY: Semantic Type; nTotal: Sum of all n's (PLC and SLCs).

Figure 3

Diagrammatic representation of the distribution of clinical domains in MeSH terms among EC forms. MeSH: Medical Subject Headings.

Figure 4

Graph showing the cumulative total frequencies among the 58 aggregated set of LP in terms of nTotal, i.e. after manual analysis and combining of all n values of PLC and SLCs for each laboratory concept. Analyzed laboratory concepts that have an nTotal above 50 concepts cover the steepest change of the slope of cumulative frequencies (Orange). Based on this, we have only included the first 55 analyzed concepts that have nTotal above 50. Blue bars show the nTotal of each analyzed laboratory concept.

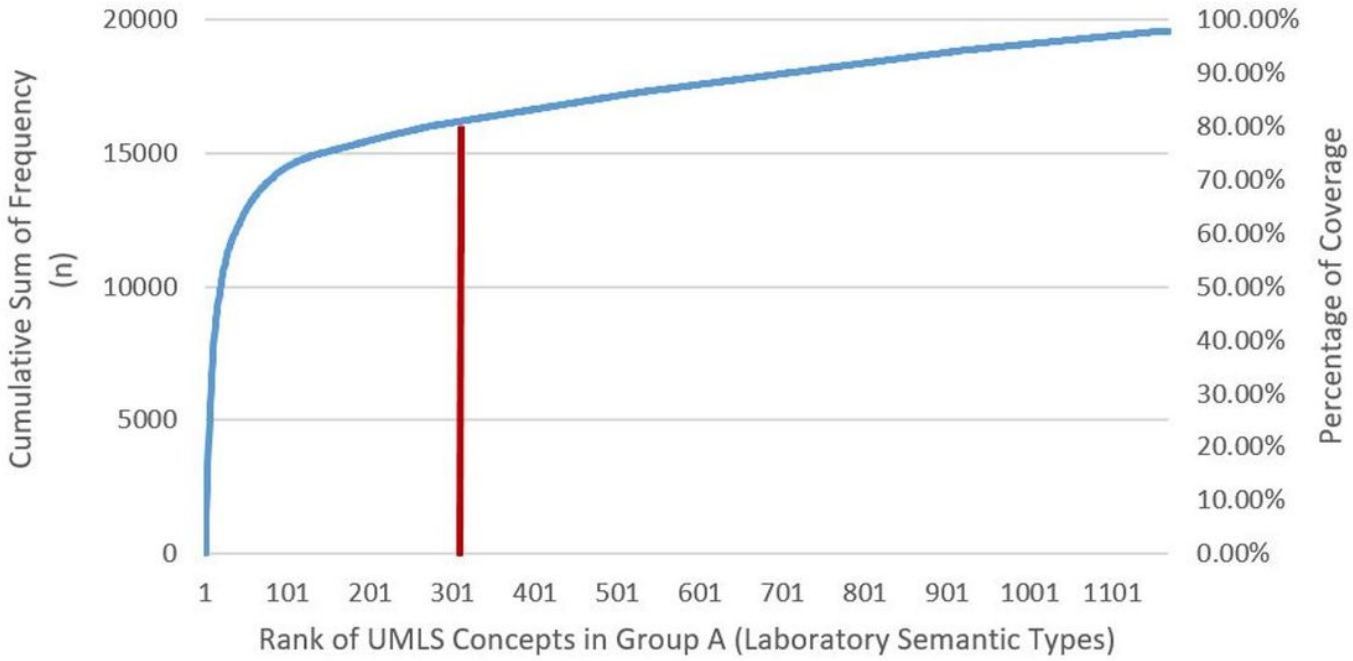

Figure 5

Plot diagram showing the coverage of laboratory concepts within the group of laboratory semantic types (Group A). 311 laboratory concepts representing our 55 LP cover 77.87% of all concept occurrences in group A. n: Frequency of Individual Concept.

Figure 6

Screenshot of the dataset on MDM portal showing available download formats e.g. ODM and FHIR.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix1FormNamesURIsNCTs.csv
- Appendix2AGroupALAB.csv
- Appendix2BGroupBNONLAB.csv
- Appendix3DETAILS.csv
- Appendix4ManualStep.pdf

- Appendix5MeSHCategories.pdf
- Appendix6SemanticTypes.pdf
- Appendix7Analysis.csv
- Appendix8ALOINCFINALDATASET.csv
- Appendix8BODM.xml
- Appendix8CFHIR.xml