Published in final edited form as:
J Biomed Inform. 2013 February ; 46(1): 33–39. doi:10.1016/j.jbi.2012.07.006.

A Human-Computer Collaborative Approach to Identifying Common Data Elements in Clinical Trial Eligibility Criteria

Zhihui Luo, PhD[*], **Riccardo Miotto, PhD**, and **Chunhua Weng, PhD**
Department of Biomedical Informatics, Columbia University, New York, NY 10032

Abstract

Objective—To identify Common Data Elements (CDEs) in eligibility criteria of multiple clinical trials studying the same disease using a human-computer collaborative approach.

Design—A set of free-text eligibility criteria from clinical trials on two representative diseases, breast cancer and cardiovascular diseases, was sampled to identify disease-specific eligibility criteria CDEs. In this proposed approach, a semantic annotator is used to recognize Unified Medical Language Systems (UMLS) terms within the eligibility criteria text. The Apriori algorithm is applied to mine frequent disease-specific UMLS terms, which are then filtered by a list of preferred UMLS semantic types, grouped by similarity based on the Dice coefficient, and, finally, manually reviewed.

Measurements—Standard precision, recall, and F-score of the CDEs recommended by the proposed approach were measured with respect to manually identified CDEs.

Results—Average precision and recall of the recommended CDEs for the two diseases were 0.823 and 0.797, respectively, leading to an average F-score of 0.810. In addition, the machine-powered CDEs covered 80% of the cardiovascular CDEs published by The American Heart Association and assigned by human experts.

Conclusion—It is feasible and effort saving to use a human-computer collaborative approach to augment domain experts for identifying disease-specific CDEs from free-text clinical trial eligibility criteria.

Keywords

Clinical Research Informatics; Clinical Trial Eligibility Criteria; Common Data Elements; Knowledge Management; Human-Computer Collaboration; Text Mining

1 Introduction

Patient recruitment is essential to successful clinical and translational research [1]. For this reason, researchers in the biomedical informatics community developed electronic screening methods that could increase the efficiency of patient recruitment [2, 3]. These methods

Corresponding Author: Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University, 622 W 168 Street, VC-5, New York, NY 10032, Tel: 212-305-3317, Fax: 212-305-3302, cw2384@columbia.edu.

[*] Division of Medical Informatics, Case Western Reserve University

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

match patients (i.e., prospective research volunteers) to the eligibility criteria for clinical trials (e.g., see the following systems: caMatch [4], Trialx [5], ResearchMatch [6]). In common nomenclature, *eligibility criteria* describe the demographic and medical characteristics that a research volunteer must possess to participate in a clinical trial, such as “body mass index <= 45 kg/m2” or “patients with insulin therapy for more than 1 week within the 3 months prior to screening”. However, due to complexities in eligibility criteria [7], this process often returns a set of clinical trials too numerous for patients to identify those for which they are eligible. For example, at the time of writing, 1,091 hypertension and 1,021 diabetes trials are actively looking for research volunteers. Merely knowing that a person has hypertension or diabetes is insufficient to include that patient in, or exclude that patient from a trial. In addition, the trial descriptions are generally provided in the form of unstructured free text. For example, ClinicalTrials.gov [8] defines only a small number of fields (e.g., “sponsor”, “study type”, “protocol location”) for registering trials using semi-structured summaries; the “Eligibility Criteria” section also remains largely unstructured[1]. Most of the existing clinical trial search engines often ask questions that are simple and limited in number (e.g., “medical conditions”, “age”, “geographic area”), a limitation that greatly reduces their specificity and thus their ability to retrieve a short list of trials that are manageable to review. Therefore, fine-grained indexing methods for clinical trial eligibility criteria are required in order to enable accurate and specific clinical trial searches [1, 9, 10].

Clinical trials studying a particular disease often employ common variables to determine patient eligibility. For example, most diabetes trials define inclusion criteria around “blood glucose level” or “hemoglobin A1c” (i.e., HbA1c), while hypertension trials typically only specify a range of values for the blood pressure. We will refer to this kind of variable as a Common Data Element (CDE). The benefits of using CDEs for clinical trial eligibility criteria in trial search are well recognized. In fact, CDEs allow users to improve the specificity of search results and minimize their need for human review. For example, in Niland’s presentation at the 2007 AMIA Annual Fall Symposium [11], a user retrieved 28 studies using the criterion “having breast cancer” alone. When breast cancer eligibility criteria CDEs were added to the query, such as “estrogen receptor status”, “progesterone receptor status”, and “cancer stage”, and a user was allowed to specify the threshold value for these variables, the number of trials retrieved decreased to seven. Other potential benefits of CDEs include their support for knowledge reuse and sharing of clinical trial eligibility criteria among investigators [12], as well as simplification of trial meta-analysis [13, 14].

CDEs for research data collection have been developed in various disease domains [15-19]. Niland *et al.*, in collaboration with CDISC, have been using an expert-driven approach to define fine-grained eligibility criteria CDEs, which enable highly specific searches of cancer trials [11]. However, such manual approaches for CDE identification generally require time-consuming discussions among experts and work for only one disease at a time [16, 19, 20]. At the time of writing, there are 116,728 trials for more than 5,000 diseases on ClinicalTrials.gov, making it impossible to manually review all the studies to identify their eligibility criteria CDEs. Therefore, the human-expert driven approach is not scalable to the large amounts of data available on this huge repository.

As a consequence, scalable approaches for eligibility criteria CDE identification are greatly needed, even if it is impossible to fully automate the task with the current technology. On the other hand, according to Friedman [21], informatics tools should augment human reasoning rather than replacing domain experts. With this design principle in mind, this article presents a semi-automated framework based on text mining that favors human-computer collaboration and assists users in the identification and refinement of eligibility

[1] http://prsinfo.clinicaltrials.gov/definitions.html

criteria CDEs. To the best of our knowledge, this is the first study exploiting text mining in the task of CDE discovery from free-text clinical trial eligibility criteria. In biomedical research, text mining has been already used to support knowledge representation and ontology generation [22]; nevertheless, current studies in these fields (e.g., the "Ontology for Biomedical Investigations (OBI) [23]", the "Ontology for Clinical Research (OCRe)" [24]) still mostly rely on manual effort for knowledge acquisition. Conversely, the proposed machine-powered solution exploits text-based knowledge acquisition and hence could potentially improve the efficiency of these projects (e.g., CDE recommendation for expert driven ontology development).

In the following, Sections 2 and 3 present the structure of the proposed approach and the experimental results, respectively. The benefits introduced by our method for eligibility criteria CDE identification and current limitations are discussed in Section 4. Lastly, Section 5 proposes future directions for the work.

2 A Semi-Automatic Approach to Derive Eligibility Criteria CDEs

Figure 1 illustrates the proposed approach to derive the eligibility criteria CDEs of any specific disease from unstructured text, which specifically refers to the free-text eligibility criteria available on ClinicalTrials.gov. As can be seen, eligibility criteria are first processed to recognize the Unified Medical Language Systems (UMLS) entities, which are then analyzed by an association-rule learning algorithm; the latter mines the CDEs, e.g., "HbA1C" associated with "diabetes" trials. These concepts are later filtered according to preferred UMLS semantic types, grouped by string similarity, and then manually reviewed as CDE candidates. Currently, the approach only identifies CDEs composed by single UMLS terms.

This study focuses on two diseases only, yet the approach is general enough to be used for any disease. In particular, we used a random sample of 1,559 breast cancer trials, with a total of 43,084 eligibility criteria, and 2,238 cardiovascular disease trials, with a total of 36,716 criteria[2]. We chose these diseases because there is a large amount of data available online at Clinicaltrials.gov as well as for the possibility of performing additional comparisons with manually annotated data (see Section 3.3 for more details).

The following sections present the components of the framework in all the most relevant details.

2.1 Semantic Annotation

The semantic annotation component tags each free-text eligibility criterion as a set of UMLS-recognizable terms. When multiple UMLS terms can be matched, the one that works best in the context of clinical trials is selected as the preferred match in order to reduce the semantic variety of the criterion itself and to favor its automatic processing. The annotator was previously described in [25] and, besides tagging the terms using an UMLS-based dictionary, it also applies a set of semantic preference rules to eliminate the inherent ambiguity in standard UMLS semantic type assignment. As an example, the UMLS term "MRI Scan" was assigned only to the semantic type "Diagnostic Procedure" (CUI-C0024485), which is more commonly used in clinical trials, discarding other available options, such as "Quantitative Concept" (CUI-C0917711). The results reported in [25]

[2] The data were obtained from Clinicaltrials.org in January 2012 using the "condition" field to filter the studies targeting these two diseases. Among all the trials, we generally retained only those containing a relevant number of eligibility criteria (e.g., we discarded the trials with zero or one criterion). In addition, preliminary results not reported here for brevity showed that the machine-powered approach needs between 1,000 and 3,000 trials to mine relevant CDEs. For this reason, we set the size of each disease corpus by sampling a number of trials falling within this interval.

shows how this approach achieves results that are at least as good as those of other solutions, such as MetaMap [26], in the context of clinical trial eligibility criteria.

2.2 Mining the CDEs

The set of UMLS terms from the sample eligibility criteria were processed by an association rule-learning algorithm to discover the CDEs associated with each specific disease. An association rule between a head **X** and a body **Y** is defined as an implication of the form $\mathbf{X} \Rightarrow \mathbf{Y}$. In our domain, **X** is the set of terms, i.e., $\mathbf{X} = \{x_1, x_2, ..., x_n\}$, that represents the patient characteristics in disease **Y** trials. For example, if **X** = {*insulin, dose*} and **Y** = {*diabetes*}, then “insulin” and “dosage of insulin” are defined as CDEs for the diabetes trials. In this domain, the support of **X**, i.e., S(**X**), is defined as the number of clinical trials containing the set **X** (i.e., the probability that **X** is present in the trials of **Y**). Therefore, the support is calculated by dividing the frequency of **X** in disease **Y** trials, by the total number of such trials.

In order to derive common association rules in the proposed approach, we used the Apriori algorithm [27], which has been already successfully employed to extract patterns within clinical variables [28-31]. The algorithm is based on a bottom-up approach where frequent patterns are extended one item at a time and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Implementation is based on breadth-first search and hash tree structure to count candidate patterns efficiently. The output is a set of rules (CDEs in this domain) that reports how often CDEs are present.

The classic version of the Apriori algorithm reduces the rule search space by exploiting the downward closure property of the support [27]. In eligibility criteria, this means that if a set of patient variables is infrequent at a specific step, then any superset built upon this set is also considered infrequent. The frequency of a set is related to the support previously defined. In our implementation, we tuned the minimum support, expressed as a probability, equal to 0.001, which implies that in the experimental sample mentioned above, a set of variables would be kept for further consideration only if the frequency was no less than three (i.e., any set occurring in less than three trials was ignored in subsequent iterations). This setup allowed us to increase the number and the diversity of the CDEs while reducing the number of false negatives (i.e., to reduce the number of true CDEs not correctly recognized). In addition, we saw empirically that a low threshold value fits the distribution of eligibility criteria, which is generally a long tail (i.e., a small number of CDEs is typically frequent, whereas others are much less so).

2.3 Filtering the CDEs

With the current setting, the Apriori algorithm returned a large number of association rules. The latter were then further filtered and grouped according to semantic similarity. In particular, we perform two post-processing steps: *semantic filtering* (see Section 2.3.1) and *similarity-based grouping* (see Section 2.3.2). One could argue that filtering and grouping could be performed using the concept of “confidence”, which is a natural measure of the importance of an association rule [32, 33]. In particular, the confidence score measures the probability of finding the rule head (i.e., **X**) of the association among all the item sets of the data (i.e., the set of eligibility criteria) containing the rule body (i.e., **Y**)[3]. However, the confidence can be affected by the size of the analyzed data, especially in presence of very unbalanced sets, i.e., those with a few rule heads that are frequent and many others that are

[3] The confidence *C* of an association rule $\mathbf{X} \Rightarrow \mathbf{Y}$ is computed using the concept of support introduced in Section 2.2; in particular, $C = S(\mathbf{X}U\mathbf{Y}) / S(\mathbf{X})$.

rare. Moreover, it does not take into account the role of the terms in the body rules; therefore, a confidence-based filter could include highly frequent terms that are not relevant CDEs (e.g., “severe”, “uncontrolled”, “clinically”). For these reasons, we decided to perform the data filtering without considering the confidence score of the association rules, but following the steps described in the next sections.

2.3.1 Semantic Filtering—*Semantic filtering* is a common technique to improve information relevance [34]. To this end, we used as our information filter a preference list of semantic types that are common in eligibility criteria. In particular, previous studies [35, 36] defined 27 semantic classes for clinical trial eligibility criteria at the sentence level, e.g., “Disease or Symptom”, “Therapy or Surgery”, “Diagnostic or Lab Results”, each containing a set of frequent UMLS semantic types. Overall we manually selected 48 semantic classes to be used as groups and filters for CDEs (see Appendix 1). For example, the semantic class “Diagnostic or Lab Results” grouped the UMLS semantic types “Laboratory Procedure” and “Laboratory or Test Results”. Therefore, the UMLS terms associated with these semantic types were designated as CDEs under the same semantic class.

2.3.2 Grouping Similar CDEs using Dice Coefficient—The list of raw CDEs returned by the Apriori algorithm was randomly ranked. As a result, similar CDEs, e.g., “urine pregnancy test”, “pregnancy test”, were often not adjacent, complicating the manual review process. Therefore, in order to return a list of CDEs ranked by some criterion of correlation, we applied the Dice Coefficient (DC) to measure the similarity between pairs of strings and created a list of CDEs ranked according to this measure.

The Dice coefficient of two strings *i* and *j* represented as bigrams is defined as

$$DC=\frac{2N_{ij}}{N_i N_j},$$

where N_{ij} is the number of bigrams in common between the two strings, and N_i and N_j are the number of bigrams for strings *i* and *j*, respectively [37].

The large number of terms extracted from the eligibility criteria made it computationally impracticable to calculate all the pair-wise similarity values. For this reason we used an iterative greedy search strategy [38], which starts by assigning the highest ranking to the term with the largest support, i.e., the “seed” term. At each iteration, the algorithm looks for the next term having the largest DC-based similarity with the current seed, adds this term to the ranking list, and makes it the new seed. The process repeats until all terms are sorted.

The result of grouping can be seen in Table 1. First, before grouping, CDEs are only tagged using the UMLS semantic types and returned in a random order (left part). Afterwards, CDEs are grouped not only by semantic type, but also by disease topic (right part, where groups are highlighted using marking symbols (i.e., *, +, ^, ~). The grouping assists manual review of the CDEs by presenting to the user semantically related concepts. In fact, in this case a user can easily see that the CDEs fall into four distinct groups: hepatic disease, heart disease, breast cancer, and function impairment.

3 Experimental Results

This section presents the results achieved by applying the proposed approach to derive CDEs for the two diseases considered, breast cancer and cardiovascular diseases.

3.1 Examples of Machine-powered Eligibility Criteria CDEs

First, we report a sample of the CDEs mined by the proposed semi-automatic approach. In particular, Table 2 shows the top five CDEs retrieved with respect to the four most frequent semantic classes for both diseases: “Disease, Symptom and Sign”, “Diagnostic or Laboratory Results”, “Pharmaceutical Substance or Drug”, and “Therapy or Surgery”. As can be seen, in most cases, our method led to pertinent CDEs. In fact, considering the breast cancer domain as an example, common variables mined by the algorithm are “chemotherapy”, “radiotherapy”, and “hormonal therapy”, which are well-known cancer therapies, as well as “IHC” (i.e., ImmunoHistoChemistry) [39] and “Platelet Count” (an indicator of the side effect of breast cancer treatments), which are frequent measures reported in related laboratory results.

3.2 Retrieval Results

We evaluated the correctness of the CDEs recommended by the proposed approach by comparing them with CDEs independently identified by the authors. Our manual identification process contained two steps: (1) list all the terms and their frequencies; and (2) manually review all these terms and retain only those that (a) occur at least three times (as the minimum support threshold defined in Section 2.2); (b) specify one patient characteristic. We used the standard information retrieval metrics precision, recall, and F-score [40]. Precision is the probability that the approach correctly retrieves a CDE. Recall is the probability that the approach retrieves a CDE that should have been retrieved. F-score is the harmonic mean of precision and recall and is a single measure of the overall retrieval performance. The results, reported in Table 3, demonstrate that the proposed approach has a high recall value, correctly identifying 80% of all the CDEs retrieved by manual review. The average precision across both diseases was 0.823, indicating that this method has a false positive rate of less than 20%. Finally, the resulting average F-score was 0.810. All the CDEs identified by the machine-powered approach and by the experts are reported in Appendix 2 and 3 for cardiovascular diseases and breast cancer, respectively.

3.3 Comparison with Expert-Defined CDEs

The American College of Cardiology Foundation (ACCF) and the American Heart Association (AHA) recently published 95 key cardiovascular disease CDEs defined by human experts [16]. For comparison, we evaluated how the cardiovascular CDEs mined by our approach were represented with respect to this standard. As Table 4 shows, the mined CDEs covered about 80% of the CDEs defined by the ACCF/AHA standard. The latter classifies all the data elements into five categories: we achieved the best coverage in the category “History and Physical Examinations” (e.g., “chest pain”, “angina grade”) with 89.2% and in “Pharmacologic Therapies” (e.g., “Aspirin and steroid”) with 85.7%. The coverage in the other categories was lower: in particular, we achieved the “Laboratory Results”, the “Diagnostic and Therapeutic Procedures”, and the “Outcomes” covered at about 73%, 67%, and 50%, respectively. It should be noted that the “Outcomes” comprised only two elements: “Death” and “Date of Death”. Our system correctly identified the former.

The proposed approach also identified some CDEs that were not formally defined by the ACCF/AHA standard. For example, Table 5 shows all the medication CDEs that the machine-powered approach found with a frequency greater than 10 that were ***not*** defined in the ACCF/AHA standard; many of these are common treatments for cardiovascular diseases. Therefore, the machine-powered approach might have the potential to augment human-based CDEs with additional knowledge. Lastly, Appendix 4 lists the comparison between the machine-recommended CDEs and the CDEs published by AHA.

4 Discussion

This section provides an analysis on the strengths (Section 4.1) and limitations (Section 4.2) of the proposed approach.

4.1 Advantages over Related Approaches to CDE Development

The human expert-based CDE development process consists of six steps: (1) reviewing data sources and existing standards; (2) generating CDE candidates; (3) prioritizing CDEs; (4) defining attributes and valid values for each CDE; (5) convening group discussions to achieve consensus definitions; and (6) eliciting peer reviews and applying for board approvals. Each step requires significant time from multidisciplinary domain experts, from several weeks to several months [15, 16, 18, 20]. The machine-powered approach can reduce the time required for domain experts on steps 1, 2, 3, and 5.

First, the existing manual CDE selection method heavily relies on domain experts in identifying CDE candidates. For example, in the ACCF/AHA standard, an informatics committee identified a preliminary set of candidate CDEs through face-to-face and conference call meetings, as well as email correspondence. This is a labor-intensive process and, in fact, the experts spent more than 6 months merely refining and vetting a list of candidate CDEs. Because our method recommends CDEs through text mining, it significantly reduces the need to manually review large amount of text.

Moreover, human-based approaches lack an objective evaluation regarding the representativeness of the CDEs within the data. For example, in order to prioritize CDEs of Atrial Fibrillation [41], the ACC/AHA committee spent several months on manually reviewing literature and trials data source. They classified CDEs by frequency into only three very broad categories: “high”, “medium”, and “low”. Conversely, the machine-powered approach provides a statistical evaluation about the role of CDEs among the various trials, which can be used in many ways. For example, the frequency information may inform trial eligibility criteria authors about the uses of CDEs in the past and their prevalence in a particular disease. It also may help researchers to develop fine-grained standards for prioritizing CDEs and reducing the manual effort needed to rank them.

Manually reviewing the CDEs also requires searching medical terminologies and existing data standards. In our approach, the CDE candidates are automatically annotated with UMLS concept ID and semantic types, and the term list is sorted to allow the user fast browsing and generalization. As a result, domain experts can start directly from the proposed CDE candidates to select and formally define CDE attributes. Therefore, the machine-powered approach has the potential to reduce the time required for clinical experts to reach a consensus.

Finally, due to the limitations of expertise specialization and time constraints, it is difficult to develop CDE candidates across a large number of different diseases. This is the main reason that existing CDE development efforts focus on one disease at a time. This is a significant barrier to achieving one of the major goals of CDE development, the support of large-scale data aggregation for research and discovery. In contrast, the machine-powered approach attempts to surmount this by making use of an unsupervised machine learning algorithm and, consequently, of a general solution. In fact, in order to work with other diseases, the approach requires only the disease name (or code) and source of information to automatically retrieve the corresponding clinical trials, parse the eligibility criteria section, and return the CDEs.

4.2 Limitations

At this point, there are two major limitations about this approach. First, some false positives were due to verbs (e.g., “schedule”, “repair”). However, they do not describe patient characteristics and should not be classified as CDEs. We hypothesize that the use of part-of-speech analysis [42] may be able to filter this type of false positive.

Second, multi-term CDEs such as “[patient’s] age at diagnosis” and “[patient’s] age at death” contain multiple UMLS terms. At this point, our system is able to identify “age” as an eligibility criterion CDE but cannot capture CDEs including multiple UMLS terms and, therefore, cannot distinguish “age of diagnosis” and “age of death”. In this case, we still rely on domain experts to group UMLS terms to form more meaningful CDEs. However, this aspect of design is impractical, and therefore temporary, because it reduces the scalability of the approach, especially with respect to the number of diseases that can be processed.

5 Conclusion

This paper contributes a scalable human-computer collaborative approach to CDE identification, which combines the use of an unsupervised machine learning algorithm (i.e., association rule-learning) with UMLS to process the free-text eligibility criteria associated with the clinical trials of a specific disease. The accuracy achieved in this initial study using two diseases, i.e., breast cancer and cardiovascular diseases, and ClinicalTrials.gov as data source is satisfactory and promising. The approach is only semi-automated, so that domain experts are still required for reviewing, filtering, and enriching the recommended CDEs. On the other hand, the machine-recommended CDEs reduce the workload of the experts, who no longer have to manually parse a large number of clinical research documents.

Future studies can focus on the following aspects. First, the semantic annotator plays a key role in CDE identification and requires improvements to accommodate the continual data updates on ClinicalTrials.gov. Second, the approach should be extended to identify multi-term CDEs to improve the overall machine performance. Third, organizing and formatting the CDEs (e.g., structuring them into ontologies) would enable interoperability. Last, this initial study focused only on CDE identification, not on CDE definition. The latter involves CDE attribute specification and feature discovery and is an important and advanced research topic on its own for semi-automated CDE discovery.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This study was sponsored under NLM grants **R01LM009886** and **R01LM010815**, and CTSA award **UL1 RR024156**. Its contents are solely the responsibility of the authors and do not necessarily represent the official view of the NIH.

7 References

1. Weng C, et al. Formal representation of Eligibility Criteria: A Literature Review. Journal of Biomedical Informatics. 2010; 43(3):451–467. [PubMed: 20034594]
2. Embi PJ, et al. Effect of a Clinical Trial Alert System on Physician Participation in Trial Recruitment. Arch Intern Med. 2005; 165(19):2272–2277. [PubMed: 16246994]
3. Thadani SR, et al. Electronic Screening Improves Efficiency in Clinical Trial Recruitment. Journal of the American Medical Informatics Association. 2009; 16(6):869–873. [PubMed: 19717797]
4. Hogarth M, et al. Introduction: caMatch a Patient Centered Clinical Trials Matching System. 2009

5. Trialx. 2/18/2012]; Available from: http://trialx.com/

6. Harris PA, et al. ResearchMatch: A National Registry to Recruit Volunteers for Clinical Research. Academic Medicine. 2012; 87(1):66–73. [PubMed: 22104055]

7. Ross, J., et al. AMIA Summit on Clinical Research Informatics. San Francisco, California: 2010. Analysis of Eligibility Criteria Complexity in Clinical Trials; p. 46-50.

8. NLM. About ClinicalTrials.gov. 2010. ClinicalTrials.govClinicalTrials.govAvailable from: http://www.clinicaltrials.gov/ct2/info/about

9. Weng C, et al. EliXR: An Approach to Eligibility Criteria Extraction and Representation. Journal of the American Medical Informatics Association (JAMIA). 2011; 18:i116–i124. [PubMed: 21807647]

10. Tu SW, et al. A practical method for transforming free-text eligibility criteria into computable criteria. Journal of Biomedical Informatics. 2011; 44(2):239–250. [PubMed: 20851207]

11. Niland, J. ASPIRE: Agreement on Standardized Protocol Inclusion Requirements for Eligibility. Mar 13. 2007 2012]; Available from: http://crisummit2010.amia.org/files/symposium2008/S14_Niland.pdf

12. Gennari JH, Sklar D, Silva J. Cross-tool communication: from protocol authoring to eligibility determination. Proc AMIA Symp. 2001:199–203. [PubMed: 11825180]

13. Sim I, Olasov B, Carini S. An ontology of randomized controlled trials for evidence-based practice: content specification and evaluation using the competency decomposition method. Journal of Biomedical Informatics. 2004; 37(2):108–119. [PubMed: 15120657]

14. Sim, I.; Olasov, B.; Carini, S. The Trial Bank System: Capturing Randomized Trials for Evidence-Based Medicine; Proceedings AMIA Annual Symposium; 2007; p. 1073

15. Cannon CP, et al. American College of Cardiology key data elements and definitions for measuring the clinical management and outcomes of patients with acute coronary syndromes. J Am Coll Cardiol. 2001; 38(7):2114–2130. [PubMed: 11738323]

16. Members WC, et al. ACCF/AHA 2011 Key Data Elements and Definitions of a Base Cardiovascular Vocabulary for Electronic Health Records. Circulation. 2011; 124(1):103–123. [PubMed: 21646493]

17. NINDS and NIH. The use of CDEs is expected to reduce study start-up times and accelerate analysis of data. 2012. http://www.commondataelements.ninds.nih.gov/

18. Patel A, et al. The development of common data elements for a multi-institute prostate cancer tissue bank: The Cooperative Prostate Cancer Tissue Resource (CPCTR) experience. BMC Cancer. 2005; 5(1):108. [PubMed: 16111498]

19. Nadkarni PM, Brandt CA. The Common Data Elements for cancer research: remarks on functions and structure. Methods of Information in Medicine. 2006; 45(6):594–601. [PubMed: 17149500]

20. Meredith Nahm AW, McCourt Brian, Pieper Karen, Honeycutt Emily, Hamilton Carol D. Harrington Robert A. Diefenbach Jane, Kisler Bron, Walker Mead, Ed Hammond W. Standardising clinical data elements. International Journal of Functional Informatics and Personalised Medicine. 2010; 3(4):314–341.

21. Friedman CP. A "Fundamental Theorem" of Biomedical Informatics. Journal of the American Medical Informatics Association. 2009; 16(2):169–170. [PubMed: 19074294]

22. Spasic I, et al. Text mining and ontologies in biomedicine: making sense of raw text. Brief Bioinform. 2005; 6(3):239–51. [PubMed: 16212772]

23. Brinkman RR, et al. Modeling biomedical experimental processes with OBI. J Biomed Semantics. 2010; 1(Suppl 1):S7. [PubMed: 20626927]

24. Samson, W.; Tu, SC.; Alan, Rector; Peter, Maccallum; Igor, Toujilov; Steve, Harris; Ida, Sim. OCRe: An Ontology of Clinical Research. 2009. Available from: http://bioportal.bioontology.org/ontologies/1076

25. Luo, Z., et al. AMIA Summit on Clinical Research Informatics. San Francisco, California: 2010. Corpus-based Approach to Creating a Semantic Lexicon for Clinical Research Eligibility Criteria from UMLS; p. 26-31.

26. Aronson, AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program; Proceedings AMIA Annual Symposium; 2001; p. 17-21.

27. Agrawal, R.; Srikant, R. Fast Algorithms for Mining Association Rules in Large Databases; Proceedings of the 20th International Conference on Very Large Data Bases; Morgan Kaufmann Publishers Inc.. 1994; p. 487-499.
28. Stilou S, et al. Mining association rules from clinical databases: an intelligent diagnostic process in healthcare. Medinfo. 2001; 10(Pt 2):1399–403.
29. Ma, L., et al. A Framework for Infection Control Surveillance Using Association Rules; AMIA Annual Symposium Proceedings; Vancouver, British Columbia, Canada. 2003; p. 410-414.
30. Nahar J, et al. Significant Cancer Prevention Factor Extraction: An Association Rule Discovery Approach. Journal of Medical Systems. 2011; 35(3):353–367. [PubMed: 20703554]
31. Karaolis, M., et al. Association rule analysis for the assessment of the risk of coronary heart events. Engineering in Medicine and Biology Society, 2009. EMBC 2009; Annual International Conference of the IEEE; 2009; p. 6238-6241.
32. Brossette SE, et al. Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. Journal of the American Medical Informatics Association. 1998; 5(4):373–381. [PubMed: 9670134]
33. Chen, J., et al. Representing Association Classification Rules Mined from Health Data; Proc. of 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems; 2005; p. 1225-1231.
34. Foltz PW. Using latent semantic indexing for information filtering. SIGOIS Bull. 1990; 11(2-3): 40–47.
35. Weng, C.; Luo, Z. Dynamic Categorization of Clinical Research Eligibility Criteria; Proc of AMIA Fall Symp; 2010; p. 306
36. Luo, Z.; Johnson, SB.; Weng, C. Semi-Automatic Induction of Semantic Classes from Free-Text Clinical Research Eligibility Criteria Using UMLS; American Medical Informatics Association Annual Symposium; Washington, DC. 2010; p. 487-491.
37. Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology. 1945; 26(3):297–302.
38. Cormen, TH. Introduction to algorithms. 3rd ed. MIT Press; Cambridge, Mass.: 2009. p. xixp. 1292
39. Harvey JM, et al. Estrogen Receptor Status by Immunohistochemistry Is Superior to the Ligand-Binding Assay for Predicting Response to Adjuvant Endocrine Therapy in Breast Cancer. Journal of Clinical Oncology. 1999; 17(5):1474. [PubMed: 10334533]
40. Manning, CD.; Raghavan, P.; Schütze, H. Introduction to information retrieval. Vol. xxi. Cambridge University Press; New York: 2008. p. 482
41. McNamara RL, et al. ACC/AHA key data elements and definitions for measuring the clinical management and outcomes of patients with Atrial Fibrillation: A report of the American College of Cardiology/American Heart Association Task Force on Clinical Data Standards (Writing Commitee to Develop Data Standards on Atrial Fibrillation). J Am Coll Cardiol. 2004; 44(2):475–95. [PubMed: 15261958]
42. Marcus MP, Santorini B, Marcinkiewicz MA. Building a Large Annotated Corpus of English: The Penn Treebank. Computational Linguistics. 1993; 19:313–330.

- A human-computer collaboration method is proposed to augment domain experts in CDE identification
- Machine-assisted CDE identification achieves acceptable accuracy
- We demonstrated the feasibility of using this method to identify about 86% of CDEs published by The American Heart Association
- UMLS plays an important role in filtering out terms with irrelevant semantic types

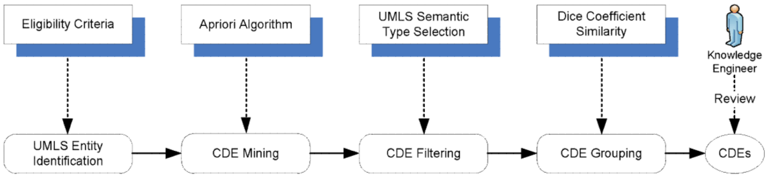

Figure 1.
A human-computer collaborative framework for identifying eligibility criteria CDEs (round-cornered boxes indicate procedures, while rectangle boxes indicate external resources or algorithms; solid arrows show workflow, and dotted arrows indicate information flow).

Table 1

Eligibility criteria CDEs as presented to the user by the Apriori algorithm (left) and after Dice Coefficient-based (DC) grouping (right). In the tables: DSNY = “Disease_or_Syndrome”; NEOP = “Neoplastic_Process”; (*, +, A, ~) mark groups of semantically related data

Type	Apriori Algorithm List
DSNY	coronary heart disease (+)
DSNY	hepatic disease (*)
NEOP	invasive carcinoma of the breast (~)
DSNY	hepatitis c (*)
NEOP	breast adenocarcinoma (~)
DSNY	hearing impairment (^)
DSNY	heart disease (+)
DSNY	cognitive impairment (^)
DSNY	hepatic cirrhosis (*)
NEOP	breast carcinoma (~)
DSNY	hepatitis b (*)
DSNY	heart attack (+)
DSNY	coronary artery disease (+)
DSNY	visual impairment (^)
NEOP	recurrent breast cancer (~)
NEOP	invasive breast cancer (~)
NEOP	stage iv breast cancer (~)
DSNY	renal impairment (^)

Type	DC-based Grouped List
DSNY	hepatitis b (*)
DSNY	hepatitis c (*)
DSNY	hepatic cirrhosis (*)
DSNY	hepatic disease (*)
DSNY	heart disease (+)
DSNY	coronary heart disease (+)
DSNY	coronary artery disease (+)
DSNY	heart attack (+)
DSNY	renal impairment (^)
DSNY	visual impairment (^)
DSNY	hearing impairment (^)
DSNY	cognitive impairment (^)
NEOP	breast adenocarcinoma (~)
NEOP	breast carcinoma (~)
NEOP	invasive carcinoma of the breast (~)

Type	DC-based Grouped List
NEOP	invasive breast cancer (~)
NEOP	stage iv breast cancer (~)
NEOP	recurrent breast cancer (~)

Table 2

The top five CDEs in the four most frequent semantic classes for breast cancer and cardiovascular diseases

Target Disease	UMLS Concept ID	Top 5 CDEs	Frequency
Therapy or Surgery			
Breast Cancer Trials	C0392920	chemotherapy	323
	C0034619	radiotherapy	98
	C0279025	hormonal_therapy	80
	C0034619	radiation_therapy	55
	C0175795	oral_medication	54
Cardiovascular Disease Trials	C1532338	pci (percutaneous coronary intervention)	747
	C0581603	revascularization	350
	C0010055	cabg (coronary artery bypass surgery)	330
	C0011946	dialysis	234
	C0162589	icd (implantable cardioverter defibrillator)	224
Diagnostic or Lab Results			
Breast Cancer Trials	C0021044	ihc (immune_histo_chemistry)	1194
	C0032181	platelet_count	776
	C0373595	creatinine_clearance	739
	C0428772	lvef (left ventricular ejection fraction)	604
	C0201976	serum_creatinine	441
Cardiovascular Disease Trials	C0428772	lvef (left_ventricular_ejection_fraction)	811
	C0201976	serum_creatinine	585
	C0032181	platelet_count	166
	C0027051	stemi (myocardial_infarction)	143
	C0302353	serum_potassium	104
Medication			
Breast Cancer Trials	C0002059	alkaline_phosphatase	696
	C0728747	trastuzumab	208
	C0039286	tamoxifen	85
	C0215136	taxane	61
	C0014939	estrogen	49
Cardiovascular Disease Trials	C0360714	statin	240
	C0003195	antiarrhythmic	110
	C0004057	aspirin	102
	C0070166	clopidogrel	100
	C0144576	paclitaxel	100
Disease, Symptom and Sign			
Breast Cancer Trials	C0006142	breast_cancer	1873
	C2939420	metastatic_disease	681

Target Disease	UMLS Concept ID	Top 5 CDEs	Frequency
	C0278488	metastatic_breast_cancer	316
	C0494165	liver_metastases	205
	C0858252	adenocarcinoma_of_the_breast	108
Cardiovascular Disease Trials	C0027051	myocardial_infarction	1706
	C0020538	hypertension	1552
	C0018801	heart_failure	1061
	C0002962	angina	897
	C0022116	ischemia	799

Table 3

Retrieval results in terms of precision (P), recall (R), and F-score for the CDE identification in the two disease domains, i.e., breast cancer, cardiovascular diseases. In the table: TP = “true positives”, FN = “false negatives”, TN = “true negatives”, FP = “false positives”

Diseases	Clinical Trials Count	Criteria Sentences (Inclusion)	Machine Powered CDEs	Human Based CDEs	TP	FN	TN	FP	P	R	F-score
Breast Cancer	849	8,616	808	829	677	152	2630	131	0.838	0.817	0.827
Cardiovascular Diseases	976	8,285	641	668	519	149	2094	124	0.807	0.777	0.792
Average									**0.823**	**0.797**	**0.810**

Table 4

Coverage of the machine-powered CDEs with respect to the ACCF/AHA standard for the cardiovascular diseases. In the table, coverage is defined as Machine / (ACCF/AHA)

CDE Source	Examination	Laboratory	Diagnosis	Pharmacologic	Outcome	Total
ACCF/AHA	37	11	24	21	2	95
Machine	33	8	16	18	1	76
Coverage	89.2%	72.7%	66.7%	85.7%	50.0%	**80.0%**

Table 5

List of medication CDEs that are not listed in the ACCF/AHA standard but are correctly found by the machine-powered approach. In the table, CUI = "Concept Unique Identifier"

UMLS CUI	Medication CDEs	Frequency
C0070166	clopidogrel	100
C0144576	paclitaxel	100
C0019134	heparin	79
C0003195	antiarrhythmic drug	74
C0040207	ticlopidine	70
C0043031	warfarin	52
C0699493	luminal	43
C0521942	angiotensin ii receptor antagonist	37
C0600437	nitric oxide donors	37
C0002598	amiodarone	19
C0012963	dobutamine	19
C0001443	adenosine	19
C0017887	nitroglycerin	16
C0541315	everolimus	16
C0003364	antihypertensive	13
C0003280	anticoagulant	12
C0017725	glucose	12
C0012265	digoxin	11
C0001617	corticosteroids	10
C0001480	atp	10