Journal of Biomedical Informatics 46 (2013) 805–813

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

# Analysis of eligibility criteria representation in industry-standard clinical trial protocols

Sanmitra Bhattacharya [a,*,1], Michael N. Cantor [b]

[a] Department of Computer Science, The University of Iowa, 14 MacLean Hall, Iowa City, IA 52242, United States
[b] Pfizer Inc., 235 E 42nd Street, New York, NY 10017, United States

ARTICLE INFO

*Article history:*
Received 5 December 2012
Accepted 3 June 2013
Available online 12 June 2013

*Keywords:*
Clinical trials
Information retrieval
Natural language processing
Controlled vocabulary
Eligibility determination

ABSTRACT

Previous research on standardization of eligibility criteria and its feasibility has traditionally been conducted on clinical trial protocols from ClinicalTrials.gov (CT). The portability and use of such standardization for full-text industry-standard protocols has not been studied in-depth. Towards this end, in this study we first compare the representation characteristics and textual complexity of a set of Pfizer's internal full-text protocols to their corresponding entries in CT. Next, we identify clusters of similar criteria sentences from both full-text and CT protocols and outline methods for standardized representation of eligibility criteria. We also study the distribution of eligibility criteria in full-text and CT protocols with respect to pre-defined semantic classes used for eligibility criteria classification. We find that in comparison to full-text protocols, CT protocols are not only more condensed but also convey less information. We also find no correlation between the variations in word-counts of the ClinicalTrials.gov and full-text protocols. While we identify 65 and 103 clusters of inclusion and exclusion criteria from full text protocols, our methods found only 36 and 63 corresponding clusters from CT protocols. For both the full-text and CT protocols we are able to identify 'templates' for standardized representations with full-text standardization being more challenging of the two. In our exploration of the semantic class distributions we find that the majority of the inclusion criteria from both full-text and CT protocols belong to the semantic class "Diagnostic and Lab Results" while "Disease, Sign or Symptom" forms the majority for exclusion criteria. Overall, we show that developing a template set of eligibility criteria for clinical trials, specifically in their full-text form, is feasible and could lead to more efficient clinical trial protocol design.



## 1. Introduction

Clinical trials are an intrinsic part of the medical research and drug development process of most pharmaceutical and biotechnology companies. Policy makers and governmental organizations are also naturally interested in evaluating the efficacy, accuracy and safety of drugs trials that could potentially affect millions of people around the world. Eligibility criteria in clinical trials are a set of requirements that a patient or participant must meet to be eligible for inclusion in a study. From the perspective of a study sponsor, these requirements ensure that all participants in a cohort satisfy some general criteria and thus give a higher confidence in predicting possible outcome of an intervention.

Eligibility criteria are usually expressed in human-readable free-text which is easily comprehensible to patient, public and researchers alike. However this form of representation of eligibility criteria makes it challenging for computable and standardized representation. Currently, there are no data or terminology standards for representing or authoring eligibility criteria in a standard format [1–3]. Given the plethora of applications of eligibility criteria, ranging from criteria reuse to patient matching from Electronic Medical Records (EMR) [4,5] it is of great importance to address the problem of computable knowledge-based representation for eligibility criteria. The primary motivation of our study is to determine the feasibility of creating a set of standard representations of eligibility criteria that would be applicable to a broad set of clinical trials. Uniformly represented eligibility criteria can facilitate the process of identification and merging of similar patient populations across studies for the purpose of patient recruitment. Consequently, this can reduce not only the time spent in performing trials that have already been conducted under similar conditions but can also help in reducing the expenses associated with participant recruitment substantially. Secondary outcomes of interest would be easier encoding of eligibility criteria to find patients via EMRs, faster and accurate authoring of eligibility criteria, and higher quality protocols. For example, a standardized representation of an eligibility criterion could be linked to a specific, standard set of

* Corresponding author. Tel.: +1 319 335 0713.
*E-mail addresses:* sanmitra-bhattacharya@uiowa.edu (S. Bhattacharya), Michael.Cantor@pfizer.com (M.N. Cantor).

[1] This study was conducted while the author was interning at Pfizer, Inc.

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.06.001

ICD-9 codes that could then be used as filters to identify patients across EMR systems.

We begin by comparing the textual and characteristic differences of industry-standard full-text protocols to corresponding protocols from ClinicalTrials.gov (CT) [6], a registry of clinical studies from around the world. We then explore methods for deriving computable and standardized representation of eligibility criteria, in the form of templates, from a set of full-text and CT protocols used in Pfizer's pain medication-related studies.

Our contributions in this paper are two-fold. First, we explore the nature of representation of eligibility criteria from industry-standard full-text protocols and compare their characteristics to corresponding CT eligibility criteria. Second, we propose a novel method for standardized representation of eligibility criteria (using sentence similarity and clustering strategies) in the form of "templates". To the best of our knowledge, this is the first study in the domain of standardized representation of eligibility criteria that deals with in-depth analysis of eligibility criteria characteristics in their full-text form. Most related research deal with considerably simpler and concise eligibility criteria from CT.

## 2. Related research

Computable clinical trial protocols and corresponding eligibility criteria representations have been studied extensively in the past two decades. Studies have been conducted to identify a set of common data elements that can be used for developing standard protocol representation [7]. There have been attempts to use natural language processing for parsing eligibility criteria statements to extract generic query patterns for eligibility criteria representation [8,9]. Research has also been focused on the identification of Unified Medical Language System (UMLS) - based [10] semantic classes for criteria statements [11,12]. The complexity of eligibility criteria representation has also been studied quantitatively with significant proportion of criteria being judged semantically complex [13].

There have been extensive studies in computer-based and formal eligibility criteria knowledge representations. A CDISC-sponsored project called ASPIRE [14] aims to provide formal representation of a core set of eligibility criteria and also provides a set of data elements which can be used for searching and filtering protocols. The Eligibility Rule Grammar and Ontology (ERGO) [15], uses an information model, composed of noun phrases, expressions and criteria, to provide a general syntax for representing eligibility criteria. The EliXR system [3] provides a semi-automated data-driven approach for semantic representation of eligibility criteria. It uses an integrated semantic processing framework based on UMLS for eliciting semantic role labels that can be used for annotating eligibility criteria. The Standards-Based Active Guideline Environment (SAGE) [16] provides a set of structured and standard terminologies for encoding computable guidelines into structured templates. The Clinical Research Filtered Query (CRFQ) Project [17] provides a standardization of criteria using various semantic parameters like demographic data, disease data, etc. The use cases of these systems vary from filtering trials satisfying particular conditions to the identification of patients for a protocol.

Previous research [18] has demonstrated the benefit of using a set of disease-categorized protocols for designing efficient clinical trial authoring tools. Significant research has also been conducted in developing decision support systems for clinical trials. Expert systems like the protocol inspection and critiquing tool (PICASSO) [19] support critiquing of clinical trial protocols and can be used to standardize new protocols. Knowledge-based decision support systems like Design-a-Trial (DaT) [20] help efficient creation of rigorous protocols documents for designed trials. The ontology-based system, TrialWiz, [21] designed to alleviate the complexity of the protocol encoding process can be used for easy authoring of clinical trial protocols. More advanced authoring tools that facilitate collaboration of protocol authors from different backgrounds have also been proposed [22]. Other than these applications, several open-source (e.g. OpenClinica [23]) and proprietary (e.g. Cytel [24], Medidata [25]) software have also been designed for assisting or automating the clinical trial protocol design process.

Although several of the above mentioned tools and applications have been developed for standardized and computable representation of eligibility criteria few can deal with the complexity of eligibility criteria as presented in full-text protocols. Most of these applications [13–15] are either semi-automated or designed using CT protocols. They still require intense manual involvement and lack flexibility for accommodating complex eligibility criteria from full-text protocols. While these tools may be capable of generating computable and knowledge-based representation of basic eligibility criteria, their utility, usability and adoptability to eligibility criteria from industry-standard protocols have not been tested rigorously. We also note that the use of automated tools for criteria standardization has very limited uptake in the industry as few of them cater to the complexities of full-text protocols. Therefore, we perform a detailed comparison of eligibility criteria representation from CT and Pfizer's clinical trial protocols and propose a novel method which can be used for industry-standard standardization. In contrast to several of the above mentioned tools, our template-based standardization approach caters to criteria reuse, which is a major objective in most pharmaceutical companies [26].

## 3. Materials and methods

In this paper we present an analysis of eligibility criteria from full-text and CT protocols across 3 dimensions. First, we compare the representation characteristics and textual complexity of eligibility criteria from full-text and their corresponding CT protocols. Second, we perform a semantic class-based comparison of these two forms of eligibility criteria representation. Finally, we generate templates based on a novel method for industry-standard full-text protocol standardization.

### 3.1. Data

We selected a set of 32 full-text clinical trial protocols in the domain of Pfizer's pain-related drug research designed between the years 2002 and 2009. In a majority of these studies, the primary objective was to evaluate the efficacy, safety or tolerability of the drugs in various patient groups under different conditions for pains related to diabetic neuropathy, total-knee arthroplasty, fibromyalgia, osteoarthritis, etc. We used the study identifier of the full-text protocols to retrieve the 32 corresponding XML-formatted protocols from CT, which currently houses over 120,000 clinical trial protocols [6]. Organizations that sponsor or conduct clinical trials are required to submit study information to a clinical trial registry like CT if they plan to publish the findings in a major journal. We wanted to compare eligibility criteria from Pfizer's full-text clinical trial protocols with the corresponding protocols retrieved from CT to assess the characteristic differences between the representation of CT and full-text. In essence, this will inform us of the complexity and processing overhead in terms of computational methods used for various studies on eligibility criteria (such as standardized representation).

### 3.2. Comparison of representation characteristics and textual complexity of full-text eligibility criteria vs. CT eligibility criteria

We first compared the representation characteristics of the eligibility statements of the full-text and CT protocols from our data-