Journal of Biomedical Informatics 54 (2015) 241–255

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

# Visual aggregate analysis of eligibility features of clinical trials

Zhe He [a], Simona Carini [b], Ida Sim [b], Chunhua Weng [a,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, USA
[b] Division of General Internal Medicine, University of California, San Francisco, San Francisco, CA 94143, USA

ARTICLE INFO

*Article history:*
Received 29 August 2014
Accepted 12 January 2015
Available online 20 January 2015

*Keywords:*
Clinical trial
Patient selection
Selection bias
Knowledge management

ABSTRACT

*Objective:* To develop a method for profiling the collective populations targeted for recruitment by multiple clinical studies addressing the same medical condition using one eligibility feature each time.
*Methods:* Using a previously published database COMPACT as the backend, we designed a scalable method for visual aggregate analysis of clinical trial eligibility features. This method consists of four modules for eligibility feature frequency analysis, query builder, distribution analysis, and visualization, respectively. This method is capable of analyzing (1) frequently used qualitative and quantitative features for recruiting subjects for a selected medical condition, (2) distribution of study enrollment on consecutive value points or value intervals of each quantitative feature, and (3) distribution of studies on the boundary values, permissible value ranges, and value range widths of each feature. All analysis results were visualized using Google Charts API. Five recruited potential users assessed the usefulness of this method for identifying common patterns in any selected eligibility feature for clinical trial participant selection.
*Results:* We implemented this method as a Web-based analytical system called VITTA (Visual Analysis Tool of Clinical Study Target Populations). We illustrated the functionality of VITTA using two sample queries involving quantitative features BMI and HbA1c for conditions "hypertension" and "Type 2 diabetes", respectively. The recruited potential users rated the user-perceived usefulness of VITTA with an average score of 86.4/100.
*Conclusions:* We contributed a novel aggregate analysis method to enable the interrogation of common patterns in quantitative eligibility criteria and the collective target populations of multiple related clinical studies. A larger-scale study is warranted to formally assess the usefulness of VITTA among clinical investigators and sponsors in various therapeutic areas.



## 1. Introduction

Well-designed clinical study protocols are essential for generating high-quality medical evidence [1]. However, studies are often criticized for lacking generalizability, or external validity [2–16]. Because population representativeness is an important aspect of clinical research generalizability, study designers should justify the tradeoffs between internal validity and external validity that arise from their choices of eligibility criteria. Biased or overly restrictive eligibility criteria may (1) exclude patients who may need or benefit from the research [4,5], and (2) lead to an overestimate of the efficacy of an intervention [10]. For example, according to Schmidt et al. [12], almost none of their analyzed studies on secondary prevention of cardiovascular events justified the applied exclusion criteria, which excluded 21–97% of the female target population. Similarly, Zimmerman et al. [15] reported that approximately 32–47% of patients with major depressive disorder would have been excluded by two most commonly used cutoff values of the Hamilton Rating Scale for Depression in antidepressant efficacy trials, i.e., 18 and 20.

When designing clinical studies, investigators often reuse eligibility criteria from previous protocols of related studies. One of our previously published papers also discovered that many clinical studies, especially those on the same medical condition, use similar or identical eligibility criteria [17]. Therefore, we hypothesize that the generalizability issue might be not only at the level of individual studies, but also at the community level in the entire clinical trial enterprise. Unlike prior work that looks at the generalizability of one study at a time, we are motivated to assess the collective generalizability by uncovering collective design patterns for participant selection among multiple related clinical trials. Unfortunately, at present there is no method or tool for making such

* Corresponding author at: Department of Biomedical Informatics, Columbia University, 622 W 168th Street, VC-5, New York, NY 10032, USA.
*E-mail address:* cw2384@cumc.columbia.edu (C. Weng).

http://dx.doi.org/10.1016/j.jbi.2015.01.005
1532-0464/

design biases transparent or help investigate such biases. Echoing this need, recently the National Center for Advancing Translational Sciences (NCATS) responded to the Institute of Medicine's review of the Clinical and Translational Science Awards (CTSA) Program in the United States and identified "lack of a knowledge base for all types of interventions at the extremes of age as well as within special populations" as one of the weaknesses of the current translational science enterprise [18]. To help bridge this gap, a computable repository of eligibility features of clinical trials is needed to analyze the characteristics of the target populations on a large scale [19].

The study and result registry ClinicalTrials.gov [20] created by the National Library of Medicine is a valuable public data source. Since September 27, 2007, all United States-based clinical trials of FDA-regulated drugs, biological products, or devices have been mandated to be registered in ClinicalTrials.gov [21]. As of March 18, 2014, 163,285 clinical studies conducted in more than 180 countries were registered in ClinialTrials.gov. Study summaries are stored in a semi-structured format in the registry, i.e., study descriptors such as title, phase, and location are organized in structured fields. The eligibility criteria are usually organized as paragraphs of free-text or as bullet lists.

The ClinicalTrials.gov is a preferred resource to be transformed into a computable repository of reusable knowledge of clinical trial designs. However, there is little published work on building a computable repository from study summaries on ClinicalTrials.gov. Tasneem et al. developed the Aggregate Analysis of ClinicalTrials.gov (AACT) database as a publicly accessible analysis dataset derived from ClinicalTrials.gov [22]. Using AACT data, clinical trials in various domains have been systematically analyzed, e.g., infectious diseases [23], oncology [24], and diabetes [25], to name a few. AACT allows selection and aggregation of trials by study descriptors, such as study status, phase, and intervention type, but not by fine-grained clinical characteristics of the target population. As studies often limit eligibility to permissible ranges of quantitative features as age, BMI, HbA1c, and blood glucose level [26], investigators or policy makers may be interested in analyzing such quantitative features across studies addressing the same medical condition, with questions like "*what is the range of BMI values that are permitted across interventional studies on Type 2 diabetes*?" However, as most of the eligibility criteria are in unstructured text, it remains difficult to support these analyses in a programmatic, accurate and scalable way. Hence, to date, there is a paucity of analyses on the quantitative eligibility features of target populations of existing studies, and consequently a lack of capacity to optimize the eligibility criteria definition for future clinical studies based on past studies.

We have developed methods for parsing eligibility features from free-text eligibility criteria [17,27–41] and the derived frequent eligibility features across ClinicalTrials.gov study summaries have produced promising results for searching and indexing studies [29], probing disease relatedness [30], and clustering studies with similar eligibility criteria [17]. Enabled by these techniques, we have created a database of discrete clinical trial eligibility features extracted from ClinicalTrials.gov called COMPACT (Commonalities in Target Populations of Clinical Trials) [42], which allows users to flexibly query sets of clinical studies (e.g., Type 2 diabetes studies) on their shared eligibility features (e.g., HbA1c or BMI) and attributes (e.g., allowed value range for HbA1c or BMI). In addition, we have developed a distribution-based method for profiling clinical trial target populations across sets of studies [43]. Meanwhile, as one of the state-of-the-art methods for discovering knowledge from Big Data [44,45], interactive visual query interfaces can be employed to further support flexible profiling of target populations of sets of clinical studies and to investigate the generalizability of these studies. It has been used for tasks similar to profiling target populations, such as visualizing alternative disease progression paths for a group of patients similar to a query patient [46], and for visual analysis of clinical event patterns through a combination of a graphical query interface, pattern mining and visualization techniques [47]. Therefore, we enhanced our COMPACT database of study summaries with visualization of the distributions of sets of clinical studies along any single quantitative eligibility feature. To the best of our knowledge, this effort represents one of the earliest attempts to perform aggregate analyses of clinical trial eligibility criteria design patterns. Fig. 1 illustrates the design of the methodology framework, which integrates text mining, data warehousing, and data visual analytics for rich information made available by ClinicalTrials.gov. This pipeline can help clinical trial designers more easily understand collective design patterns in clinical trial eligibility criteria across multiple related clinical trial studies. On this basis, our system can increase the transparency of hidden eligibility criteria design biases at the clinical research community level. Our system supports flexible study selection using multiple study descriptors, such as study type, study design, intervention type, phase, condition, gender, and age range. We hypothesized that our method could identify understudied population subgroups whose value ranges for certain quantitative eligibility features were systematically excluded or overly researched according to analyses of eligibility criteria specifications. Our preliminary user evaluation confirmed this hypothesis and the value of our method for improving the transparency of clinical trial participant selection decisions.

The remainder of the paper is organized as follows. Section 2 first describes the visual aggregate analysis system of eligibility features of clinical trials and how a user interacts with it, and then delineates the methods used to develop and evaluate the system. In Section 3, we use Type 2 diabetes and hypertension as example conditions to illustrate the functionalities of the system. We also present the results of a preliminary evaluation with a convenience sample of five potential users of the system. Finally, we discuss the implication and the limitations of this work in Section 4 and draw conclusions in Section 5.

## 2. Methods

Previously, we introduced a novel database called COMPACT, which stores metadata and parsed eligibility criteria of study summaries in ClinicalTrials.gov [42]. It supports retrieval of readily analyzable eligibility features, quantitative or qualitative, from sets of studies. On this basis, we designed an interactive visual analysis system to aggregate target populations of sets of clinical studies. The potential users of this system include clinical investigators, study sponsors and policy makers. Table 1 presents the glossary of terms that are frequently used in this paper.

Our system enables a user to select a medical condition, one of the quantitative eligibility features frequently seen in studies on that medical condition and other additional study descriptors (e.g., study type, study design, intervention type) to perform five analyses: (1) distribution of number of studies over consecutive value points or non-overlapping value intervals within user-specified value range of the selected quantitative eligibility feature (e.g., over each 0.5% of HbA1c); (2) distribution of enrollment over those value points or value intervals; (3) distribution of number of studies over boundary values (e.g., lower bound of HbA1c as 7.0%); (4) distribution of number of studies over permissible value ranges (e.g., BMI between 15 and 25 kg/m$^2$); and (5) distribution of number of studies over value range widths (e.g., the value range width for HbA1c between 7% and 10% is $10 - 7 = 3$). Fig. 2 illustrates the comparison between information provided by the ClinicalTrials.gov and our system called VITTA (Visual Analysis Tool of Clini-