A Cohort Discovery and Exploration System for Clinical Trial Accrual

Andrew Wen, MS[1], Huan He, PhD[1], Sunyang Fu, PhD[1], Sijia Liu, PhD[1], Kurt Miller, MS[1], Robert Gehrke[2], Carmen Vodislav[2], Michael Lin [2], Kathryn Cook[2], David Strauss MBA[4], Dania Helgeson MS[4], Thomas Kingsley, MD, MPH[3,4], Alex Ryu, MD[4], Hongfang Liu, PhD[1]

[1] Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
[2] Center for Clinical and Translational Science, Mayo Clinic, Rochester, MN, USA
[3] Department of Quantitative Health Sciences, Mayo Clinic, Rochester, MN, USA
[4] Department of Medicine, Mayo Clinic, Rochester, MN, USA

Introduction

Patient recruitment for randomized clinical trials (RCTs) has continued to be a serious challenge in healthcare for the past decade[1]. A recent study reported that around 80% of trials fail to recruit enough patients to meet the initial enrollment target [2], leading to reduced statistical power and premature discontinuation of the RCT, consequently threatening the certainty of any results. Therefore, many commercial solutions have been proposed and many pilot studies have been conducted to facilitate rapid patient recruitment, which has shown that one of the potential approaches is to reuse the data from electronic health record (EHR) systems to increase recruitment efficiency [3].

Leveraging EHR data for patient recruitment remains, however, challenging. First, as no EHR system modules are yet directly dedicated to supporting patient assessment for RCTs, research coordinators and physicians must perform the time-consuming task of manually searching and browsing clinical facts in the EHR system to retrieve relevant information. Secondly, evaluating the eligibility criteria for each patient not only requires extensive scientific knowledge but also takes much effort to check all the relevant datasets. Existing tools or methods can be helpful in reducing the workload, but the data and outputs in each tool are not directly connected to each other, limiting their practical effectiveness and efficiency.

To address these challenges, we propose a cohort discovery and exploration system to streamline the eligibility assessment workflow based on open-source techniques of information retrieval, human-computer interaction, and natural language processing (NLP).

Methods

Although many NLP-based techniques have been proposed to assess criteria automatically, the scientific expertise of research coordinators is still strongly demanded to evaluate patient eligibility due to the complexity of inclusion/exclusion criteria in clinical trials. Therefore, we propose a user-centric approach to assist the user in cohort discovery and exploration. As shown in Figure 1a, the user can edit eligibility criteria and the query for assessment, browse the cohort extracted from the EHR system, and identify potential recruits with aggregated clinical facts and scores, which are supported by multiple techniques of information retrieval, human-computer interaction, and NLP.

To implement this integrative approach, we designed a full-stack system architecture to combine the modules from different fields. As shown in Figure 1b, our proposed system consists of four major components: web-based frontend applications, a web service middleware component, a horizontally scalable backend, and a data warehouse.

First, the web-based frontend application provides the user interface (UIs) for conducting the major tasks related to patient recruitment. Among these frontend applications, the eligibility reviewer provides the main features for reviewing the inclusion/exclusion criteria for a candidate, which is designed based on human-computer interaction principles and developed based on web frontend techniques to visualize the criteria and facts for decision making.

Secondly, the web service middleware component provides evidence services, cohort services, and project services through a RESTful API to support both the frontend and backend applications. The middleware is designed in a microservice architecture approach, which enables better system scalability and reusability.

Thirdly, the backend implements modules for cohort retrieval and patient assessment, such as the query translator that converts UMLS codes to local vocabulary, term scoring that ranks the relevance of documents to a given criterion, and storage adapters that unify the data representation. Said backend is implemented using Apache Beam following the MapReduce paradigm, allowing for horizontal scaling across a variety of cluster computing architectures.

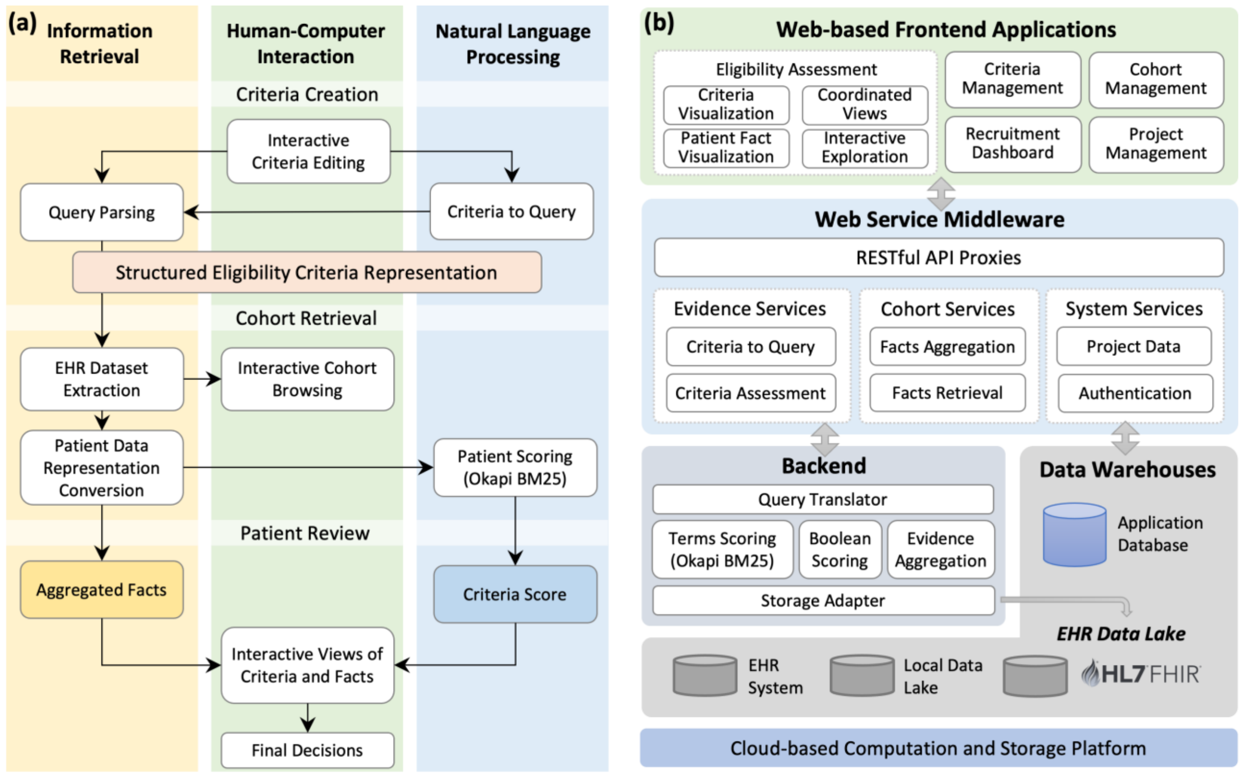

Figure 1. (a) Our proposed approach for eligibility criteria assessment, which integrates the techniques from information retrieval, human-computer interaction, and natural language processing to support the major tasks for clinical trial accrual. (b) The architecture of our proposed cohort discovery and exploration system, which consists of web-based applications, web service middleware, back, and data warehouse.

Lastly, the output of the backend modules and the user-generated datasets, such as user determinations and algorithm judgments, are saved in the data warehouse for further data analysis. This architecture can be implemented in both cloud computing environments and local standalone servers to meet the needs of scalability and data security.

Discussion and Future Work

During the development of our system, we found that as the EHRs are mainly used for healthcare and clinical needs, relevant data is neither always structured nor presented in a manner compatible with the clinical trial eligibility criteria. Several studies also show that the completeness of EHR data may not be suitable for the clinical trial accrual process [4]. In addition, it is not an easy task for researchers or physicians to “translate” the eligibility criteria into queries that can be processed by an EHR system. As a result, an informatician is strongly required in the clinical trial accrual to bridge the gap between the fields of clinical research and information system. In terms of this requirement, the role and abilities of the informatician should be taken into the system design consideration.

As a next step, we are going to integrate several NLP methods into our system to improve its ability in criteria translation and assessment. Secondly, we plan to conduct case studies on several clinical trials to validate the effectiveness of our proposed system with our collaborators. The source code and more technical details will be released on our GitHub repository: https://github.com/OHNLP/CohortDiscoveryTool.

References

1. Brøgger-Mikkelsen M, Ali Z, Zibert JR, Andersen AD, Thomsen SF. Online Patient Recruitment in Clinical Trials: Systematic Review and Meta-Analysis. Journal of Medical Internet Research. 2020 Nov 4;22(11):e22179.
2. Johnson O. An evidence-based approach to conducting clinical trial feasibility assessments. Clinical Investigation. 2015 May;5(5):491–9.
3. Köpcke F, Trinczek B, Majeed RW, Schreiweis B, Wenk J, Leusch T, et al. Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. BMC Medical Informatics and Decision Making. 2013 Mar 21;13(1):37.
4. Lai YS, Afseth JD. A review of the impact of utilising electronic medical records for clinical research recruitment. Clinical Trials. 2019 Apr 1;16(2):194–203.