Brief Communication

The COVID-19 Trial Finder

Yingcheng Sun1,†, Alex Butler1,2,†, Fengyang Lin3, Hao Liu1, Latoya A. Stewart4, Jae Hyun Kim1, Betina Ross S. Idnay5,6, Qingyin Ge3, Xinyi Wei3, Cong Liu1, Chi Yuan1, and Chunhua Weng1

1Department of Biomedical Informatics, Columbia University, New York, New York, USA, 2Department of Medicine, Columbia University, New York, New York, USA, 3Department of Statistics, Columbia University, New York, New York, USA, 4Vagelos College of Physicians and Surgeons, Columbia University, New York, New York, USA, 5Department of Neurology, Columbia University, New York, New York, USA, and 6School of Nursing, Columbia University, New York, New York, USA

†Authors contributed equally.

Corresponding Author: Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University, 622 W 168 ST, PH-20 room 407, New York, NY 10032, USA; chunhua@columbia.edu

Received 13 September 2020; Editorial Decision 13 November 2020; Accepted 19 November 2020

ABSTRACT

Clinical trials are the gold standard for generating reliable medical evidence. The biggest bottleneck in clinical trials is recruitment. To facilitate recruitment, tools for patient search of relevant clinical trials have been developed, but users often suffer from information overload. With nearly 700 coronavirus disease 2019 (COVID-19) trials conducted in the United States as of August 2020, it is imperative to enable rapid recruitment to these studies. The COVID-19 Trial Finder was designed to facilitate patient-centered search of COVID-19 trials, first by location and radius distance from trial sites, and then by brief, dynamically generated medical questions to allow users to prescreen their eligibility for nearby COVID-19 trials with minimum human computer interaction. A simulation study using 20 publicly available patient case reports demonstrates its precision and effectiveness.

Key words: clinical trial, eligibility criteria, COVID-19, information filtering, questionnaire, web application

INTRODUCTION

Patient-to-trial matching remains a critical bottleneck in clinical research, largely due to the free-text format of clinical trial information,1 particularly eligibility criteria that are indispensable for screening patient eligibility and yet not amenable to even simple computation.2 Existing clinical trial search systems are either keyword-based or questionnaire-based.3 Keyword-based search engines, such as ClinicalTrials.gov, FindMeCure.com, Janssen Global Trial Finder,4 or ResearchMatch,5 require users to search for trials using keywords, which tends to impose challenges for query formulation and generate information overload.6 Static questionnaire systems, such as Fox Trial Finder,7 filter out irrelevant trials by asking users to answer a long list of preselected questions, which can be laborious and are not user-friendly.

The coronavirus disease 2019 (COVID-19) pandemic is one of the greatest challenges modern medicine has faced. As of August 2020, there have been more than 6 million confirmed cases and 180 000 reported deaths in the United States, with few approved treatments.8,9 In response to the COVID-19 emergency, clinical trial research assessing the efficacy and safety of COVID-19 treatments are being created at an unprecedented rate. As of August 31, 2020, well over 3100 clinical trials have been registered in ClinicalTrials.gov, the largest clinical trial registry in the world. The need for rapid and accessible trial search tools has never been more apparent than now.

In this article, we describe an open-source semantic search engine for COVID-19 clinical trials conducted in the United States called the COVID-19 Trial Finder by extending our previously published method for using dynamically generated questionnaires for enabling
efficient clinical trial search. This is an interactive COVID-19 trial search engine that enables minimized, dynamic questionnaire generation in response to user provided answers in real time. It is powered by a regularly updated machine-readable dataset for all the COVID-19 trials in the United States. It is also enhanced with a Web-based visualization of the geographic distribution of COVID-19 trials in the United States to enable friendly user navigation with the trial space. By facilitating search for appropriate COVID-19 trials in specific geographic areas, the system enables research volunteers to perform self-screening using the eligibility criteria of these COVID-19 trials. Further, it allows clinical trialists to assess the landscape of COVID-19 trials by eligibility criteria and geographical locations in order to identify collaboration opportunities for similar COVID-19 studies and improve trial response corresponding to evolving case surges.

The system is accessible online (https://covidtrialx.dbmi.columbia.edu), as well as its source code (https://github.com/WengLab-InformaticsResearch/COVID19-TrialFinder). We evaluated the system on 20 published COVID-19 case reports and demonstrated its precision and efficiency.

**MATERIALS AND METHODS**

The COVID-19 Trial Finder consists of 2 modules, for trial indexing and trial retrieval, respectively. The trial indexing module works offline to extract entities and attributes from eligibility criteria text and to create a trial index using semantic tags, which are the extracted terms mapped to standardized clinical concepts. The retrieval module dynamically generates medical questions and iteratively filters out trials based on user answers until a sufficiently shortened list of trials is generated. Figure 1 shows the system architecture.

Clinical trial eligibility criteria exist largely as free text, so they must be formalized to a machine-readable syntax to allow for semantic trial retrieval. In the trial indexing module, COVID-19-related trials are acquired from ClinicalTrials.gov by querying all the trials indexed with “COVID-19” being their condition. Using a semi-automated method, their eligibility criteria are structured and formatted using the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) first by an automated tool Criteria2Query, and then verified and corrected manually as needed by medical domain experts (A.B., J.K., L.S.) to overcome the limitations in Criteria2Query. With international memberships in the terminology sets, the OMOP CDM provides a comprehensive standard vocabulary for representing clinical concepts commonly available inpatient data. Take an example in our system, the phrase “shortness of breath” in the criteria text will first be extracted by Criteria2Query through its named entity recognition module and mapped to the concept “dyspnea” as the semantic tag through the entity normalization module, which will then be used to index associated trials. We provided a dataset with this article, including 581 COVID-19 trials annotated by 10 223 semantic tags, with 17.6 tags per trial on average. After removing duplicates, 1811 distinct tags are used for trial index. Supplementary Table 1 lists a subset of the dataset. Newly registered COVID-19 trials are updated weekly to the database, and the trial index list is updated accordingly.

Entities from the following 5 domains are used for question generation: condition, device, drug, measurement, and procedure. Domain-specific question templates are provided in Supplementary Table 2. For example, “Have you ever been diagnosed with (condition concept)?” is the question template for condition entities. The templates are not exhaustive to cover all possible questions but are designed to triage trials by building on prior knowledge of common eligibility criteria.

The trial retrieval module interacts with users and asks criteria-related questions to facilitate eligibility determination. Users first enter their location (eg, zip code) and then select a study type (eg, interventional, observational). Next, 5 most frequently used criteria about current age, high-risk status (eg, hospital worker), COVID-19 status, current hospitalization or intensive care unit admission, and pregnancy status are formulated as “standard questions” and posed to all users. These criteria are frequently used across all COVID-19 trials and thus serve as an important participant stratification step, so that they are posed together on a single page instead of via dynamically generated questions. Afterward, an eligibility criterion with the maximum information gain (ie, with the highest entropy) is selected and rendered into question using the corresponding templates. Based on the user’s answer to the presented question, the
most ineligible trials are filtered out. This process iterates, each time narrowing down the trial candidates pool and visualizing the recruiting sites of the remaining eligible trials on an interactive map, until the user reaches a short list of trials.

Four main Web interfaces are shown in Figure 2. On the index page, users can specify the geographical area to search for recruiting sites by inputting a zip code and an adjustable radius in section 1. Section 2 shows a collapsible section containing advanced search options such as trial type, recruiting status, and keyword search. Five “standard questions” are expected to be answered in section 3, and users can skip them by clicking the “skip questions” button in section 4 and enter the dynamic questionnaire page. Users can select the question type and answer one question at a time in section 5, and the candidate trial list is dynamically updated and shown in section 6. The answered questions are recorded in section 7, where users can return to previous questions to make updates. After clicking the “Show Eligible Trials” button, users will be able to visualize them geographically. Eligible trials are listed by their titles in section 8, and all the recruiting sites within the user delimited area are marked as small green icons on the map section 9. The embedded map is powered by Google Maps application programming interface. Users can select any trial to review additional details such as study type, description, contact information, and location(s) in section 10, and its recruiting sites within the geographic area as specified by the user in section 1 will be highlighted and pinpointed on the map as well. Additionally, participants interested in learning more about the study will be able to access a link to ClinicalTrials.gov. The initial version of the COVID-19 Trial Finder was released in May, and more than 690 page visits from 20 countries were recorded by the end of August 2020, including 615 visits from 40 states in the United States, according to the report of Google Analytics.

EXPERIMENT

We evaluated the effectiveness of the system by assessing its precision in identifying appropriate trials for users. We selected 20 patient cases in the U.S. from COVID-19 case reports curated by LitCOVID, with consideration for diversity in location, age, sex, and comorbidities, and run simulations on our system based on these patient cases. The detailed information of the 20 cases can be found in the Supplementary Table 3. For each case report, the zip code was based on the corresponding author’s address stated in the case report. We set the default radius as 100 miles to ensure adequate coverage of available trials. The 5 standard questions and multiple dynamic questions were answered based on the patient profile and reported symptoms. We continuously answered the dynamic questions until no more questions could be generated and the system prompted a review of the returned trial list. An example of the question-answering process is shown in Supplementary Table 4.

We then manually reviewed each identified clinical trial in the final list to confirm its relevance to the user query by examining the inclusion and exclusion criteria available at ClinicalTrials.gov based on the patient case report (ie, the clinical trial was deemed relevant if the user met all inclusion criteria and no exclusion criteria). If there was no information in the case report to determine if the patient met any of the exclusion criteria, we considered the clinical trial relevant because eligibility could be further determined by the clinical research staff once the user initiates contact for possible participation. Finally, we calculated the individual search precision for each patient case by comparing the number of identified trials in the final list to the number of trials manually confirmed relevant. The averaged precision for the 20 user cases was calculated to indicate the system precision.

Next, we evaluated the efficiency of the system by comparing the number of trials identified at each step of the search. The percentage of trials being filtered was the number of identified trials after answering the 5 standard questions divided by the number of retrieved trials after answering dynamically generated questions.

RESULTS

Table 1 demonstrates the diversity of the user cases. Patient age ranged from 43 days to 80 years, with 25% under 30 years of age, 55% between 30 and 60 years of age, and 20% over 60 years of age. The locations were distributed across 10 states in the United States encompassing 13 different counties. On average, 14 questions were answered for each patient case report, yielding an average precision of 79.76% in finding eligible trials. Because the number of trials found by the system varied considerably for each case (eg, only 1 trial was identified for cases 3 and 20, while 34 were identified for case 10), precision was normalized by the number of trials after screening for each case. On average, 34.8% of trials were filtered out after answering 9 dynamic questions, which is consistent with the experimental results for efficiency evaluation of DQueST.8

DISCUSSION

A small number of identified trials were irrelevant as confirmed by manual review. In review, the impression encountered in finding eligible trials was largely caused by the inability to generate relevant questions. For a few criteria, no questions were asked to filter out ineligible trials, and these limitations can be summarized into 3 types: location, identity, and condition. Examples alongside the unmatched criteria and causes for the errors are described in Table 2.

The “location” limitation refers to insufficient granularity or specificity in our question template for locations. The “identity” limitation signifies insufficient specificity in the identity of the participant, such as some trials recruiting clinical therapisists instead of COVID-19-infected patients. For the “condition” limitation, extraction may be incorrect or missed so that concepts are mismatched to the terms.12 For example, the word severe can be a qualifier for a condition as opposed to be part of condition definition.

To improve the relevance and precision of the trial filtering, this system could utilize a more granular annotation model to cover more entities and attributes as well as a wider range of domain types for these annotations such as visit, person, or observation within the OMOP model. Further, additional question templates would allow an increased number of questions to be posed. Considering the tradeoff between finer granularity in the annotation model and the increase in annotation cost, we did not add more questions in this study, but future efforts can explore how to efficiently annotate more types of criteria to boost the precision of trial matching while maintaining a high level of usability and comfortable ease of access to maintain user participation.

Currently, we included only COVID-19 trials conducted in the United States in the Trial Finder application simply to keep the scope manageable for evaluation purpose and to avoid the need for engineering work on translating the system into different foreign languages. Our open-source method is available for adoption and implementation by researchers across the world. We compared the inclusion and exclusion criteria of 777 COVID-19 trials in the
United States and 2318 COVID-19 trials outside the United States registered on ClinicalTrials.gov by October 1, 2020, and found 42.3% of overlap (87.0% if not counting infrequent criteria, which are defined as criteria that appear in <10 trials). For the different criteria, they can also be indexed with standard concepts and searched by corresponded questions because the OMOP CDM...
includes international terminologies. It will be interesting and feasible to apply the Trial Finder system for non-U.S. trials in the future.

CONCLUSION
The COVID-19 Trial Finder facilitates fast search and self-eligibility screening for COVID-19 trial seekers. Despite its limitations, preliminary evaluation by emulated case reports demonstrates its precision and efficiency, showing its potential as a user-friendly COVID-19 trial search engine.

FUNDING
This work was supported by the National Library of Medicine grant R01LM009886-11 (Bridging the Semantic Gap Between Research Eligibility Criteria and Clinical Data) and National Center for Advancing Clinical and Translational Science grants UL1TR001873 and 3U24TR001579-05 (to CW).

AUTHOR CONTRIBUTIONS
YS, AB, and CW conceived the system design together. YS, AB, FL, and CL designed and implemented the system. CW supervised the design and implementation. HL, LAS, JHK, and CY contributed to the data annotation. BRSI, QG, and XW contributed to the evaluation of the system. All authors edited and approved the manuscript.

SUPPLEMENTARY MATERIAL
Supplementary material is available at Journal of the American Medical Informatics Association online.

DATA AVAILABILITY STATEMENTS
The data underlying this article are available in Dryad Digital Repository, at https://doi.org/10.5061/dryad.7bj4402z9 (https://datadryad.org/stash/share/XWwjmqkOcRkXofSvPg-XNCahhexMbE(Gf4eoa07KTFeA)).
CONFLICT OF INTEREST
None declared.

REFERENCES
1. Kang T, Zhang S, Tang Y, et al. EliIE: An open-source information extraction system for clinical trial eligibility criteria. *J Am Med Inform Assoc* 2017; 24 (6): 1062–71.
2. Sun Y, Loparo K. Information extraction from free text in clinical trials with knowledge-based distant supervision. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 1; 2019: 954–5.
3. Utami D, Barry B, Bickmore T, et al. A conversational agent-based clinical trial search engine. In: *Proceedings of the Annual Symposium on Human-Computer Interaction and Information Retrieval (HCIR)*; 2013.
4. Amy R. Global Trial Finder: why it just got easier to enroll in a Janssen clinical study, 2016. https://www.jnj.com/latest-news/global-trial-finder-why-it-just-got-easier-to-enroll-in-%20a-janssen-clinical-study Accessed December 7, 2020.
5. Pulley JM, Jerome RN, Bernard GR, et al. Connecting the public with clinical trial options: the ResearchMatch trials today tool. *J Clin Trans Sci* 2018; 2 (4): 253–7.
6. Liu C, Chi Y, Alex MB, et al. DQueST: dynamic questionnaire for search of clinical trials. *J Am Med Inform Assoc* 2019; 26 (11): 1333–43.
7. Meunier C, Jennings D, Hunter C, et al. Fox trial finder: an innovative web-based trial matching tool to facilitate clinical trial recruitment. *Neurology* Apr 2012; 78 (Meeting Abstracts 1): P02.241.
8. Wissel BD, Van Camp PJ, Kouril M, et al. An interactive online dashboard for tracking covid-19 in us counties, cities, and states in real time. *J Am Med Inform Assoc* 2020; 27 (7): 1121–5.
9. COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University (JHU). ArcGIS. Johns Hopkins University. Accessed August 31, 2020.
10. Reich C, Ryan PB, Belenkaya R, et al. OHDSI Common Data Model v6.0 Specifications. https://ohdsi.github.io/CommonDataModel/ Accessed December 7, 2020.
11. Kury F, Butler A, Yuan C, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. *Sci Data* 2020; 7 (1): 281.
12. Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. *J Am Med Inform Assoc* 2019; 26 (4): 294–305.
13. Sun Y, Butler A, Lin F, et al. Data from: The COVID-19 Trial Finder. Dryad Digital Repository. https://doi.org/10.5061/dryad.7b44j0x9 Accessed December 7, 2020.
14. Chen Q, Allot A, Lu Z. Keep up with the latest coronavirus research. *Nature* 2020; 579 (7798): 193.