AMIA
INFORMATICS PROFESSIONALS. LEADING THE WAY.

OXFORD

Research and Applications

DQueST: dynamic questionnaire for search of clinical trials

Cong Liu,[1] Chi Yuan,[1] Alex M. Butler,[1,2] Richard D Carvajal,[2] Ziran Ryan Li,[1] Casey N. Ta,[1] and Chunhua Weng[1]

[1]Department of Biomedical Informatics, Columbia University, New York, New York, USA, [2]Department of Medicine, Columbia University, New York, New York, USA

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

ABSTRACT

Objective: Information overload remains a challenge for patients seeking clinical trials. We present a novel system (DQueST) that reduces information overload for trial seekers using dynamic questionnaires.

Materials and Methods: DQueST first performs information extraction and criteria library curation. DQueST transforms criteria narratives in the *ClinicalTrials.gov* repository into a structured format, normalizes clinical entities using standard concepts, clusters related criteria, and stores the resulting curated library. DQueST then implements a real-time dynamic question generation algorithm. During user interaction, the initial search is similar to a standard search engine, and then DQueST performs real-time dynamic question generation to select criteria from the library 1 at a time by maximizing its relevance score that reflects its ability to rule out ineligible trials. DQueST dynamically updates the remaining trial set by removing ineligible trials based on user responses to corresponding questions. The process iterates until users decide to stop and begin manually reviewing the remaining trials.

Results: In simulation experiments initiated by 10 diseases, DQueST reduced information overload by filtering out 60%–80% of initial trials after 50 questions. Reviewing the generated questions against previous answers, on average, 79.7% of the questions were relevant to the queried conditions. By examining the eligibility of random samples of trials ruled out by DQueST, we estimate the accuracy of the filtering procedure is 63.7%. In a study using 5 mock patient profiles, DQueST on average retrieved trials with a 1.465 times higher density of eligible trials than an existing search engine. In a patient-centered usability evaluation, patients found DQueST useful, easy to use, and returning relevant results.

Conclusion: DQueST contributes a novel framework for transforming free-text eligibility criteria to questions and filtering out clinical trials based on user answers to questions dynamically. It promises to augment keyword-based methods to improve clinical trial search.

Key words: clinical trial search, natural language processing, common data model, interactive search, eligibility criteria

INTRODUCTION

Recruitment is a costly bottleneck in clinical research. Typical informatics approaches to recruitment largely fall into 2 categories: investigator-led patient screening or patient-centered clinical trial search. The universal adoption of electronic health records (EHR) has made it possible to use e-screening to identify eligible patients.[1] Meanwhile, the advances in the Internet technologies have also

boosted patient-centered trial search platforms.[2] Many trial search engines, such as *ClinicalTrials.gov*,[3] www.findmecure.com, and www.trialstoday.org provide patient-centered trial search services. Most of these existing trial search engines are keyword-based and retrieve relevant trials from preindexed repositories, which inevitably cause information overload in response to generic user inputs. For example, a search for "breast cancer" in ClincalTrials.gov returns more than 8000 records as of September 2018. Current strategies to reduce information overload rely on querying against structured information, such as demographic criteria, study location, and recruitment status, which is useful to a certain degree for retrieving relevant trials but falls short in retrieving eligible trials due to the underutilization of free-text eligibility criteria, resulting unmanageable false positives.

Recent technologies have used free-text eligibility criteria to facilitate trial search. For example, eTACTs[4] developed by Miotto et al[5] extracted preannotated Unified Medical Language System terms from eligibility criteria so that trials were indexed by frequent eligibility tags, with each tag's importance in trial searching indicated by a combination of font size and color in a word cloud view. When users click a tag, eTACTS filters the remaining trials by the selected tag and dynamically generates tag clouds using the tag frequency distribution in the remaining trials. While trial search benefits from an interactive information retrieval process, the semantics or the underlying meaning embedded in eligibility criteria have not been fully utilized. For example, filtering by the tag "*BMI*" alone is insufficient for matching patients to the criterion "*inclusion: BMI > 40*" because it still requires the understanding of the meaning of the expression "BMI > 40" and its comparison to patients' BMI.

A number of efforts have focused on representing the semantics of free-text eligibility criteria in a structured format to facilitate computational processing.[6–8] Both an ontology-based and template-based representation of eligibility criteria for standardizing criteria statements have been proposed. For example, using natural language processing (NLP) and a rule-based approach, EliXR[9] recognized the Unified Medical Language System concepts and their frequent combinatory patterns. Similarly, EliXR-TIME[10] and Valx[11] have been developed for temporal and numeric expression extraction and normalization, respectively. Machine learning based approaches, such as EliIE[12] and its extension Criteria2Query,[13] have been developed to identify standardized medical entities and related attributes in eligibility criteria as well.

With the advances in structured representation for free-text eligibility criteria, we see the promise of converting standardized eligibility criteria to a set of eligibility questions so that eligible trials could then be returned based on the answers to those questions. For example, a criterion "*inclusion: BMI > 40*" could be converted to a corresponding question "*What is your BMI?*" and patient eligibility could then be determined by patients' answers. Conventional questionnaires are designed for a specific trial and the data are collected and stored using clinical trial management systems such as Research Electronic Data Capture application (REDCap).[14] Those static questionnaires tend to be long (up to a few hundred questions) due to their static nature and, hence, involve tedious efforts (up to hours) for patients to answer.[15] To cover the criteria of thousands of trials related to a single condition, the number of questions in a generic static questionnaire can exceed the capacity of patients to complete in a reasonable time frame.

Dynamic questionnaires have the potential to reduce the data collection burden by asking only questions still applicable after previous answers.[16] In this study, we developed a system for dynamic questionnaire generation and implemented an interactive question-answering module to optimize the trial search efficiency. Our goal is to generate the most informative and accurate questions from a standardized criteria library to retrieve eligible trials in the most efficient fashion. The system aims to ask as few questions as possible to identify eligible trials with high accuracy. To the best of our knowledge, this is the first effort to automatically and dynamically select the most informative eligibility criteria and generate questions to rapidly reduce the search space for clinical trial. A web application (http://impact2.dbmi.columbia.edu/dquest-flask/) was developed to demonstrate this framework. We evaluated its performance by using both simulation and a scenario-based studies and we evaluated the usability of this system by letting a small cohort of patients at New York-Presbyterian/Columbia University Irving Medical Center test the tool and provide feedback.

MATERIALS AND METHODS

The framework of DQueST

DQueST consists of 2 modules (Figure 1). Module 1 works offline to perform information retrieval and criteria library curation. The eligibility criteria texts were extracted from the eligibility criteria section of all clinical trials in *ClinicalTrials.gov* as of August, 2018 and then parsed into standardized formats using an extension of our previously published open-source pipelines.[12,13] Extracted clinical entities are mapped to standard concepts and stored in a criteria library. In order to accelerate the trial search, criteria clustering is conducted to compress the size of the library. Module 2 implements a real-time dynamic question generation algorithm. The process starts with an initial search, similar to a standard search engine, and a working library is created for the following iterative criteria-based filtering. The iterative process then starts by considering all criteria in the working library. At each step, a criterion is selected from this criteria library according to its relevancy score, which reflects the question's ability to filter out ineligible trials, and converted to a corresponding question. After patients answer the corresponding questions, the candidate pool (of trials) and the working library are updated by filtering out ineligible trials and deleting the previously selected criterion, respectively. The process iterates until users decide to stop and begin manually reviewing the remaining trials.

Module 1: information retrieval and criteria library curation

Information extraction

Free-text eligibility criteria for 252 330 trials archived in the *ClinicalTrials.gov* repository (as of August, 2018) were processed and transformed to a standardized representation using an extension of previously published work.[12,13] Free-text eligibility criteria were then split into paragraphs by line breaks and further split into segments utilizing a sentence splitting method from Stanford CoreNLP.[17] For each criteria segment, the output is a clinical entity with its attributes including domain, temporal constraint, numeric expression and negation status. A sequence labeling method, conditional random fields,[18] was employed for clinical entity and its attribute recognition with the model trained from an annotated corpus based on the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).[12] The same model was also used for domain prediction. NegEx,[19] a rule-based algorithm, was then used to detect the negation status. We applied an "AND" logic
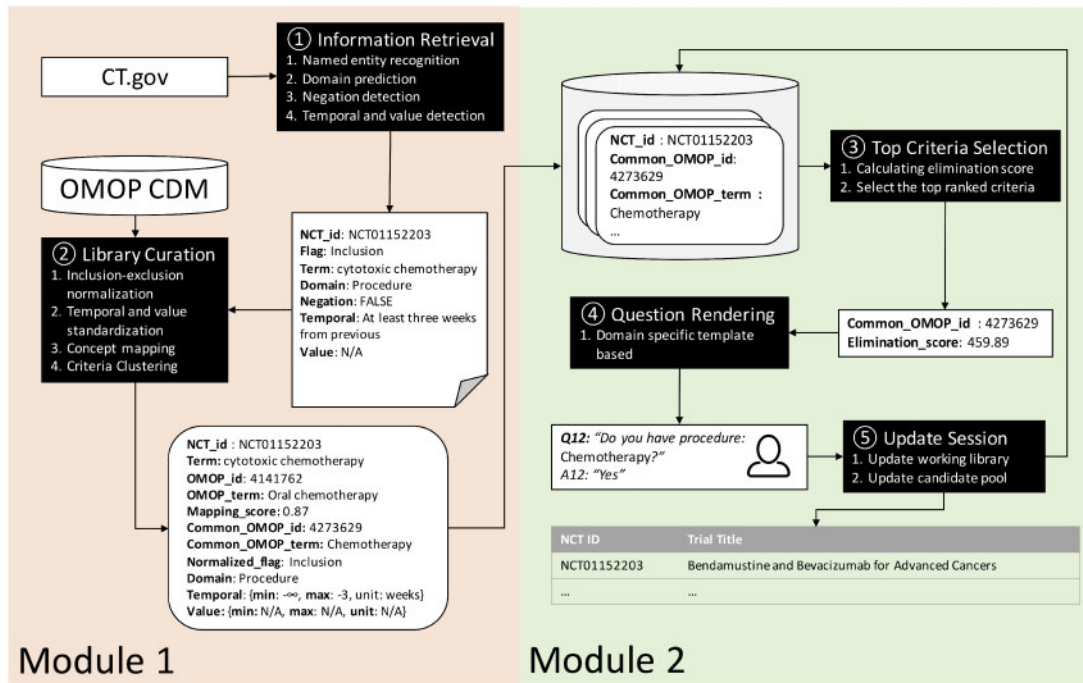
Figure 1. The pipeline architecture of the DQueST system. (A) Module 1 works offline to retrieve information from the trial repository and curate the eligibility criteria library; (B) module 2 interacts with users and generates questions dynamically.

among all clinical entities in inclusion criteria and an "OR" logic among all clinical entities in exclusion criteria.

Library curation

All negated entities were then converted into the opposite category for inclusion or exclusion. For example, "*nonmalignant origin present*" in inclusion criteria was transformed to "*malignant origin present*" in exclusion criteria. Since we have assumed "AND" logic operations for inclusion and "OR" operation for exclusion, the negation conversion will not alter the logic. Temporal constraints and numeric expressions were extracted and normalized to a minimum and maximum range using a rule-based script. The normalization for numeric values is straightforward. For temporal values, we set the time of enrollment as 0 and normalize the temporal value accordingly. For example, "*Participants with history of breast cancer in the past 5 years*" would be parsed as "*{min: -5, max: 0, unit: year}*."

All of the clinical entities extracted were then mapped to the standard concepts as defined in OMOP CDM[20] using the Usagi tool (https://github.com/OHDSI/usagi) provided by observational health data sciences and informatics. Hereafter, all mentions of "concepts" refer to "OMOP concepts." Usagi is a Java application making use of Apache's Lucene Java library for entity matching based on string similarity.[21,22] All of the unmapped or low-quality mapped entities were discarded if the mapping score (range: 0–1) was < 0.7, a threshold based on empirical heuristics. The "concept–relationship" table maintained by the OMOP CDM was used to convert synonyms and nonstandardized concepts to standardized concepts. To compress the size of the criteria library, we used a clustering strategy to merge similar or related concepts into 1 common concept (as described later).

Coverage score calculation and clustering

To calculate the coverage score for each concept, we defined $m(e, c)$ as the mapping score from an entity e to a concept c (from Usagi), and $I(e)$ is the number of trials with an entity e. We then calculated the coverage score

$$S(c) = \sum_{e:e \rightarrow c} I(e)m(e, c)$$

for each unique OMOP concept c in the library, which is the weighted summation of the mapping scores for all trials with an entity e mapping to the concept c.

We then clustered concept c' to c only if (1) $S(c) > S(c')$, (2) c is an ancestor of c' in OMOP vocabulary, (3) the maximal separation between c and c' is less than 2 in the hierarchical tree, (4) c is not in a manually curated blacklist (such as "disease," "operation," etc; more details described later), and (5) $S(c') < 20$. The concepts were sorted ascendingly according to the coverage scores and the above clustering procedure was executed from top to bottom, 1 concept each time, with the coverage score recalculated and the list resorted.

The first rule leverages the evidences from all eligibility criteria defined in clinical trials to select the most common concepts, and the second rule employs the rich hierarchical relations maintained in OMOP CDM to provide the higher-level abstraction. The third and fourth rules are implemented empirically to guarantee the quality of question approximation by limiting the degree of clustering. The fifth rule is employed to avoid clustering concepts with high mapping coverage. For example, "oral chemotherapy" is clustered as "chemotherapy" rather than "procedure" because "procedure" is a blacklisted concept and more than 2 levels of separation above "oral chemotherapy."

A "blacklist" was manually curated to identify high-level concepts not suitable for question generation. To generate this blacklist, a co-author (AB) manually reviewed the top 200 most frequently occurring concepts in each domain and noted if they were not suitable to be converted into a specific question. For example, concepts such as "Disease," "Illness," "Procedure," "Sensitivity," and "Pain" are

Table 1. Table schema of the criteria library generated after extraction and mapping

Field Name	Example	Description
ID	12345678	Unique ID for the record
NCT_ID	NCT01152203	NCT ID associated with the record
CLINICAL_ENTITY	Cytotoxic chemotherapy	Clinical term extracted from the original text
OMOP_ID	4141762	Standard OMOP concept ID mapped to the clinical entity
OMOP_TERM	Oral chemotherapy	Standard OMOP concept name
MAPPING_SCORE	0.87	Similarity score between the mapped OMOP concept and the clinical entity
COMMON_OMOP_ID	4273629	OMOP concept ID after clustering
COMMON_OMOP_TERM	Chemotherapy	OMOP concept term after clustering
FLAG	INCLUSION	This is an inclusion or exclusion criteria
DOMAIN	Procedure	The domain of the OMOP concept
VALUE_MIN	N/A	Standardized minimum of numeric expression associated with this concept
VALUE_MAX	N/A	Standardized maximum of numeric expression associated with this concept
VALUE_UNIT	N/A	Unit of numeric expression associated with this concept
TEMPORAL_MIN	$-\infty$	Standardized minimum of temporal constraint associated with this concept
TEMPORAL_MAX	-3	Standardized maximum of temporal constraint associated with this concept
TEMPORAL_UNIT	weeks	Unit of temporal constraint associated with this concept

too general and are not considered useful in this platform's screening questions (eg, "Have you ever been diagnosed with – *Disease*?").

We did not perform concept clustering for concepts falling into the domain "measurement" due to the heterogonous interpretations, different measurements, and range of "values" associated with different lab tests. For measurement concepts, DQueST requests users to type in the actual value of the measurement (if they can provide one). Since the ancestor and descendant measurement concepts can differ in what are being measured, how things are measured using different measurement units, and/or the reference ranges being used to classify normal vs abnormal lab values, simply clustering related concepts can cause un-neglectable errors. For example, considering the criterion "[INCLUSION] platelet count < 50 000/μL," the ancestor of "platelet count" is "blood cell count." If we performed clustering for the measurement concepts, then the question will be "What is your recent blood cell count?" However, this question is ambiguous, since blood cell count can refer to different measurements (eg, red blood cells, white blood cells, platelets), thus the patient's answers for "blood cell count" cannot be used as an answer for "platelet count."

The criteria library was organized in a table structure. Table 1 summarizes the table schema. Each record is specified by a unique combination of trial and standardized concept. The example depicted in Table 1 defines an inclusion criterion in trial "NCT01152203": "*At least 3 weeks from previous cytotoxic chemotherapy.*" "*Cytotoxic chemotherapy*" was extracted and mapped to a nonstandardized OMOP concept "Oral chemotherapy (& cytotoxic)," remapped to a standard concept "Oral chemotherapy," and then further clustered to a common concept "Chemotherapy."

Module 2: Online dynamic questionnaire
Top criteria selection
DQueST always starts from an initial search, similar to any standard keyword-based search engine, and then iteratively reduces the search space by asking criteria-related questions and applying the answers to further filter the resultant trials. Notably, this module is designed to generate these criteria-related questions dynamically. The module starts from a copy of the criteria library (ie, working criteria library), and at each step, the algorithm identifies the most informative common concept c from the working criteria library based on a relevancy score. The relevancy score was calculated for all common

concepts remaining in the working library. The relevancy score $RS(c, L)$ was defined as following:

$$RS(c, L) = \sum_{c' : c' \to c} \sum_{e : e \to c'} I_L(e) m(e, c')$$

where $I_L(e)$ is the number of unique trials with entity e in the working library L, and $m(e, c')$ is the mapping score for entity e. The scores of individual entities were then summed into the score for the common concept. The relevancy score was designed to achieve 2 goals: (1) providing information to as many trials as possible in the working library, which is controlled by $I_L(e)$; and (2) high precision of the question representing the original criteria, which is controlled by $m(e)$.

Question rendering
Each selected common concept was rendered as a series of questions based on the domain category of the concept. A question series is usually composed of 2 of 3 aspects (present status, value range, or temporal constraint), and different templates were prepared for parameterizing different domains, including condition, observation, measurement, drug, and procedure (Table 2). For the measurement, the criteria are applied to the most recent measurement. The example in Table 1 would render the questions: "Have you had *Chemotherapy*?" and "If yes, what is the procedure period *(for the recent Chemotherapy)*?"

Dynamic update
A patient is considered ineligible for a trial if the patient's answer matches an exclusion criterion or does not match an inclusion criterion. A "match" is defined as fulfilling the following condition according to the question template: (1) the answer for questions regarding presence is "Yes"; (2) answer of value falls into the value range of this criteria; (3) the answer of temporal periods overlaps with the temporal range in the criteria. Patients may choose to skip questions that they do not want to answer. Ineligible trials, and all of the records related to these trials together with previously selected common concepts, are removed from the candidate pool and the working library, respectively. The top criterion is again selected from the updated working library and the process iterates until the number of unique trials in the candidate pool is reduced to an appropriate level (as decided by users).

Table 2. Question templates used for different domains

	Question Template	Answer Template
	Condition	
Present	Have you ever been diagnosed with [condition_concept]? (*required*)	yes/no/don't know
Value	N/A	N/A
Temporal	Could you provide the start and end time? (*optional*)	start_date, end_date
	Observation	
Present	Do you currently have or have you ever had/been [observation_concept]? (*required*)	yes/no/don't know
Value	N/A	N/A
Temporal	Could you provide the start and end time? (*optional*)	start_date, end_date
	Measurement	
Present	Do you know your most recent [measurement_concept]?	yes/no/don't know
Value	Please enter the value: (*required*)	value_as_number/NULL
Temporal	N/A	N/A
	Drug	
Present	Have you ever taken or received [drug_concept]? (*required*)	yes/no/don't know
Value	N/A	N/A
Temporal	Could you provide the start and end time? (*optional*)	start_date, end_date
	Procedure	
Present	When you ever undergone a(n) [procedure_concept]? (*required*)	yes/no/don't know
Value	N/A	N/A
Temporal	Could you provide the start and end time? (*optional*)	start_date, end_date

Design of patient-centered usability evaluation

To test the usability of this platform in a real-world setting, we recruited a cohort of 12 patients to use the online tool (http://impact2.dbmi.columbia.edu/dquest-flask/) in the Division of Hematology/Oncology at New York-Presbyterian/Columbia University Irving Medical Center. Patients were asked to use the DQueST Platform on an iPad Mini. They were not asked to enter their own medical information into the tool and could answer questions using arbitrary answers. Their total time spent on the tool and total number of questions answered were measured. Following the use of the tool, they were asked to complete a 10-question System Usability Survey modified from the System Usability Scale created by John Brooke in 1986.[23] Finally, some general demographic information was collected, including gender, age group, medical literacy, technologic literacy, and whether they completed the form with someone else (family member, partner, etc).

RESULTS

User interface and system implementation

A web demo is available at (http://impact2.dbmi.columbia.edu/dquest-flask/). Its source code is available at (https://github.com/stormliucong/dquest-flask). Similar to other search engines, regular keyword, demographic, geographic, and recruitment status-based searches were implemented in our web app to allow users to find trials that interest them and to reduce the search space by answering questions. For example, for most of the patients who are interested in only active trials, they can use the advanced search function provided to limit their search space. After the size of the candidate pool is reduced by the initial search, users can use the question-guided search to further filter out ineligible trials. A table of retained candidate trials displays basic trial information along with links to the *ClinicalTrials.gov* engine to provide additional details. Figure 2 shows an example question panel. When users click the *confirm* button, the table of candidates are updated given the user's input, and users can change their answers before they click the *confirm* button, which confirms the update and prompts

the next question. The procedure repeats until users are satisfied with the candidate trials.

The back end of the system is a criteria library stored in an SQL relational database. The working library is a per user/session product. Our design did not store the working library on the server side; instead it is maintained in the client side as a JSON object. The client will send the server a request with this JSON object whenever an update is requested (ie, press the confirm button), and the server will return an updated JSON object back to the client. Therefore, each session/user will have its own working libraries.

Search space compression

Table 3 shows how the search space is compressed from entities to standard concepts and further to concept clusters. On average, each unique entity is associated with 11.6 trials. About 72.5% of the entities were removed due to their low-quality mapping to the OMOP CDM standard concepts. Most of the discarded entities belonged to the measurement and drug domains. Since multiple entities can be mapped to a single concept, the index set was reduced by 77.0% via concept mapping, and the concept clustering reduced this number further by 65.5%.

The iterative dynamic questionnaire can then reduce the number of candidate trials (compress the search space) based on user answers. We selected 10 diseases associated with more than 1000 trials each in the *ClinicalTrials.gov* repository. For each disease, we evaluated the percentage of trials filtered out after answering a certain number of questions (10, 20, 30, 40, 50) dynamically prompted by the DQueST system. The answers to the questions were simulated by random selection. The experiments were repeated 100 times. In general, DQueST reduced information overload by filtering out 60%–80% of trials after asking 50 questions, with the rest of trials lacking common eligibility criteria to be easily filtered. Figure 3 demonstrates the average percent of trials filtered out after a variable number of questions are answered. The logarithm-like curve was expected since the early questions can provide more information and, thus, on average lead to a faster shrinkage in search space.

Figure 2. The DQueST user interface. (A) Trial searches can be initialized by searching for keywords, demographics, and location using the API provided by *ClinicalTrials.gov*. Users can use advanced search to restrict to only ongoing trials. (B) Question-guided interactive trial searching is followed to filtered out ineligible trials dynamically. The remaining trials in will be updated and a new question will be asked once the users click the confirm button. (C) the remaining eligible trials are shown and the users can click the link to see more details about a specific trial in the *ClinicalTrials.gov* repository. (D) The users can use the navigation panel to provide a different answer to any previous questions or simply review the remaining trials at any question stage.

Table 3. Basic statistics for the record, entity, concepts and concept cluster

Domain	Measurement	Condition	Drug	Observation	Procedure
Total number of entity occurrences	811 822	3 187 262	1 010 898	312 375	882 553
Total number of unique entities	104 334	302 426	93 815	3977	30 116
Total number of unique entities mapped to the OMOP CDM	13 665	107 763	14 314	1385	10 004
Total number of unique concepts	4547	18 684	6128	853	3581
Total number of unique concept clusters	N/A	4094	4720	642	2193

Different conditions showed various performance largely due to the difference in the total number of initially relevant trials. For example, there were approximately 2600 trials related to obesity according to the keyword-based search in *ClinicalTrials.gov*. After answering 50 questions, on average > 75% of trials were removed. This observation suggests DQueST could be more useful when the number of candidate trials is larger.

Question relevance

We optimized this system to maximize the relevancy of the questions, though evaluation of the question relevance is difficult due to the lack of well-established gold standards. Using the same set of 10 conditions described above, 1 physician and 1 biomedical informatician examined the questions and rated the relevance of the questions according to their knowledge to the query conditions. The experiment was repeated 10 times with randomized answers provided for the first 10 questions on each pass. The union of the questions generated from the first 10 rounds for each condition were then rated. On average, 79.7% of the questions were determined to be relevant to the query conditions. Among the irrelevant questions, most questions were gender specific (eg, *Patient currently pregnant*), which should be adjusted according to the patient's gender. Another type of irrelevancy is caused by the granularity of the question. For example, questions like "*Do you have illnesses*" are too general to be useful. A better handcrafted blacklist should be maintained to avoid over-clustering. Table 4 shows the top 10 questions (common concepts) appearing in the most trials. As expected, "pregnant" is the most frequently occurring criterion among all trials. A list of most frequent criterion (common concepts) for different conditions are shown in Supplementary Material File S1. Question rankings vary with different conditions. For example, the most frequently occurring criterion in obesity related trials is "Body mass index," which is the key measurement for "Obesity."

Accuracy of filtering

DQueST was designed to automatically filter out ineligible trials. We conducted an experiment in Alzheimer's disease related trials to evaluate the accuracy of the filtering. We provided a random answer
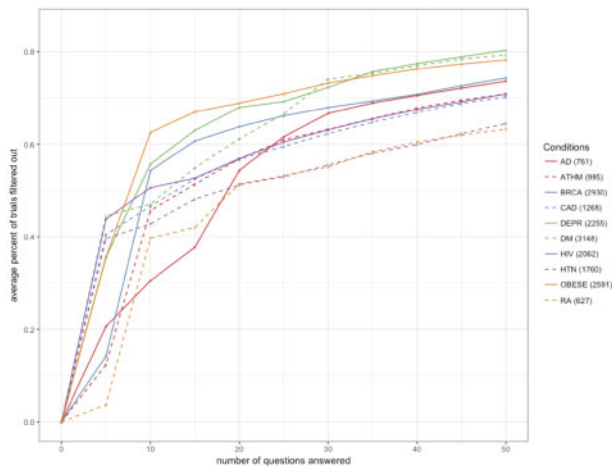
Figure 3. The average percent of trials filtered out (total number of trials) vs the number of questions answered.

Abbreviations: AD, Alzheimer's disease; ATHM, Asthma; BRCA, Breast cancer; CAD, Coronary artery disease; DEPR, Depression; DM, Diabetes mellitus; HIV, Human immunodeficiency virus infection; HTN, Hypertension; OBESE, Obesity; RA, Rheumatoid arthritis.

Table 4. The top 10 concepts identified in eligibility criteria associated with the most trials

Concept Name	Domain	Number of unique trials associated with
Patient currently pregnant	condition	76 719
Allergic reaction to substance	condition	22 832
Allergic disposition	condition	19 175
Myocardial infarction	condition	15 745
Malignant neoplastic disease	condition	14 065
Body mass index	measurement	13 333
Chemotherapy	procedure	12 080
Cerebrovascular accident	condition	10 173
Congestive heart failure	condition	8720
Blood coagulation disorder	condition	8708

for the first 10 questions in DQueST and evaluated the accuracy by verifying the eligibility of 10 randomly sampled trials (or all trials, if the number was less than 10) filtered out at each question. On average, the accuracy of filtering is 63.7%. Table 5 shows some example errors. Most of the errors (86.7%) were caused by the processing methodology of negated inclusion and exclusion criteria. That is, a "negation" event in an exclusion criterion could not simply convert to a non-negated inclusion criterion. For example, in the exclusion criteria "Pregnant women will not be eligible to participate," DQueST extracts "pregnant" as a "negated" exclusion criterion and erroneously converts it into a "pregnant" inclusion criterion. More evaluation details were shown in Supplementary Material File S2.

Eligibility of search results

DQueST was designed to find clinical trials that patients may be eligible for. Therefore, it is important to evaluate the quality of the returned trials by eligibility. We generated 5 mock patient profiles with different conditions (Supplementary Material File S3). We then asked 2 reviewers (1 biomedical informatician and 1 physician) to answer the questions generated by the system based on the profiles of the mock patients. They could stop answering questions whenever they think the change of returning trials are subtle. An example is shown in Supplementary Material File S4. Ten trials were then randomly sampled from the returned trials, and the 2 reviewers assessed the percentage of eligible trial. We compared the results against the first 10 results returned by searching the keywords in *ClinicalTrials.gov*. Table 6 summarizes the eligibility results of the 2 systems. On average, the trials returned by DQueST were 1.465 times more relevant than the trials returned by *ClincalTrials.gov*.

Scenario-based user-centered evaluation

Descriptive statistics of the real-world patients surveyed is shown in Supplementary Material File S6, and data collected from the patient surveys is described in Table 7. In short, patients surveyed were generally comfortable using DQueST, with most scores regarding "ease of use" or "tool comprehension" being "Somewhat Agree." Of note, "I would need the support of a technical person to be able to use this system," was the question that had responses closest to "Neutral,' highlighting a need to augment the user interface toward a more patient-centered target audience. Additional correlation analysis is shown in Supplementary Material File S7. All comments from the surveyed patients can be seen in Supplementary Material File S5, but the most common comment related to the tool being

Table 5. Examples of trials that are erroneously filtered out

Question	Answer	NCT_ID	Criteria and its Section	Error Reason
Have you had dementia?	Yes	NCT00948766	Exclusion Criteria: A current diagnosis of probable or possible vascular dementia.	Inappropriate granularity
Have you taken Memantine?	Yes	NCT00477659	Exclusion Criteria: memantine (Namenda) are permitted during the study	"permitted" in exclusion
Have you had patient currently pregnant?	No	NCT02958670	Inclusion Criteria: male study subjects with female partners who are pregnant	Condition for family member
Have you had epilepsy?	No	NCT00814697	Exclusion Criteria: They must not have a history of seizures or epilepsy.	"must not" in exclusion
Have you taken Donepezil?	Yes	NCT02968719	Exclusion criteria: Hypersensitivity to donepezil or piperidine	mismatched question and criteria

Table 6. Estimated percentage of remaining trials considered to be eligible for the mock patient by physician and biomedical informatician, respectively

Mock patient	System	Physician	Biomedical Informatician	Mean between Physician and Informatician
1	DQueST	20%	50%	35%
	ClinicalTrials.gov.	10%	30%	20%
2	DQueST	50%	60%	55%
	ClinicalTrials.gov.	40%	40%	40%
3	DQueST	30%	40%	35%
	ClinicalTrials.gov.	20%	30%	25%
4	DQueST	20%	40%	30%
	ClinicalTrials.gov.	20%	20%	20%
5	DQueST	60%	70%	65%
	ClinicalTrials.gov.	50%	50%	50%

Table 7. User evaluation of the usability of DQueST

	Question Mean[a]	Std. Dev.
1 I found the tool unnecessarily complex	1.92	0.90
2 I thought the tool was easy to use	4.00	0.43
3 I would need the support of a technical person to be able to use this system	2.42	1.44
4 I would imagine that most people would learn to use this tool very quickly	3.92	0.90
5 I thought the questions were clear and easy to understand	4.67	0.49
6 I thought the questions were relevant to my search	4.25	0.97
7 I am likely to use this tool to search for clinical trials	4.08	1.00
8 I found this tool more useful than other trial search tools I have used in the past	3.92	1.24
9 The number of questions to identify clinical trials was appropriate	4.50	0.52
10 How likely are you to recommend this tool to others?	9.33	0.78
# of Questions Answered	5.33	3.73
Time Spent on Questionnaire (minutes)	5.83	1.95

[a]"For question 1–9, '1' indicates 'Strongly Disagree', '5' indicates 'Strongly Agree'. For question 10, '1' indicates" No recommendation, and "10" indicates "Strongly recommended."

"tough on first attempt" and the need for a "description of [the] tool at [the] start." Following this feedback, in the future, our team will design a tutorial targeted at first-time users to familiarize them with the interface and to explain the purpose of the tool.

DISCUSSION

In this study, we present a novel framework for dynamic question-guided clinical trial search that integrates NLP techniques and OMOP CDM based criteria representations. Successes of dynamic questionnaires has been reported in other domains. For instance, ThinkGenetics[24] has developed a system, SymptomMatcher, to render phenotype questions and identify genetic causes based on patients' answers to the questions. Compared with conventional keyword-based search methods, DQueST has several unique advantages. First, by utilizing the semantic pattern embedded in eligibility criteria, it has the ability to retrieve eligible clinical trials based on patient information. For example, HbA1c measurement is a commonly used criterion (eg, *Subject has an HbA1c value between 7.0% and 10.5%*). A keyword-based search engine cannot filter out trials based on patient's HbA1c value. In contrast, DQueST could leverage a patient's response to the question "What is your HbA1c?" to reduce the search space immediately.

Second, term mismatches between index terms and query terms have been a serious obstacle to the enhancement of the retrieval performance.[25] It is difficult for patients using a keyword search to know which terms to include in order to effectively narrow down their search results. DQueST uses an interactive question-answering design where the patient's search is "guided" by a data-driven approach that identifies the most relevant questions. Third, DQueST is a personalized search engine: the questions are selected given patients' responses to the previous questions, and the overall searches are tailored according to patients' clinical profiles.

Finally, rather than competing with standard trial search engines, DQueST is instead complementary. DQueST starts from a first-round search, similar to any standard search engine, later reducing the search space iteratively by asking criteria-related questions, and applying the answers to filter out the resultant trials. By providing a dynamic questionnaire after a customized initial search, DQueST has the potential to support a wide array of use cases and stakeholders. The primary use case of DQueST is for patients searching for trials into which they can enroll, but it can also be used by other stakeholders including clinicians and investigators. For example, if a clinician or researcher wanted to test a new therapy in a specific population of patients, they could complete the questionnaire using the information corresponding to their target population and identify current as well as previous trials in that research space. This would allow them to better assess the treatment landscape as well as identify previous therapies for further review.

Curation of a good criteria library is key to the success of the system. The library consists of a set of common data elements for clinical trial eligibility criteria. Many efforts have been focused on identifying common data elements in clinical trial eligibility criteria. For example, Luo et al developed a human–computer collaborative approach to augment domain experts for identifying disease-specific common data elements. In this study, we utilized a machine learning based method to train a named entity recognition model and further mapped the entities to standard OMOP concepts. We acknowledge that the domain category of an entity and its associated attributes are predicted based on a machine learning trained model, which could be biased by the training corpus. In addition, string-similarity based concept mapping techniques could lead to errors. Alternatively, other sophisticated NLP systems such as MetaMap,[26,27] cTakes,[28] and MedLEE[29] could be used to identify the common data elements embedded in the eligible criteria. In order to maximize the power of the evolving technology in NLP,[30] we adopted a modular design for our framework and made the information extraction and criteria library curation independent modules. Module 1 works offline to feed a standardized criteria library to module 2 for question-guided trial search. In the ideal case, the eligibility criteria library could be manually (or semi-automatically) curated in a structured format to provide the most accurate representation of eligibility criteria,[6,31] and our optimization module will work independently to prompt informative questions dynamically.

In DQueST, we decided to use the OMOP CDM for the following reasons. First, OMOP is a robust data model and can be utilized by many stakeholders.[30] Therefore, by adopting OMOP, DQueST remains connected to the OMOP ecosystem and can be more easily adopted or integrated into other tools. For example, in the event that a trial recruiter is looking to identify eligible trials for a candidate, they can apply the patient's EHR data (in an OMOP data model) to DQueST to perform this screening. Secondly, the OMOP CDM is a person-centric model that accommodates different data domains typically found within observational data (demographics, visits, condition occurrences, drug exposures, procedures, and laboratory data).[20] Similarly, DQueST uses a person-centric design in which questions are derived from the concepts in various data domains. Compared to Unified Medical Language System, the OMOP data domain can lead to more meaningful questions since it is designed to collect observational data. Thirdly, our NLP module is trained on a training corpus annotated using the OMOP CDM. Therefore, solely from a technical standpoint, it is easier and more accurate for us to process the criteria by using the OMOP CDM.

The idea of question-guided interactive search is to leverage the common data elements in eligibility criteria and to prioritize the questions based on its current expected information gain. There is a trade-off between accuracy and efficiency. The higher the degree of trials connected via some common data elements in the criteria library, the faster the trial filtering process. However, increasing the connectivity degree of the common data elements requires a more aggressive concept mapping (eg, including low-quality mapping results) and clustering (eg, clustering to much higher-level concepts) strategy, which is likely to introduce higher error rates due to mapping errors and overclustering. In the most extreme cases, if we do not conduct any concept mapping or clustering, the patient will be forced to answer more questions, further slowing this filtering process. DQueST controls the granularity of the criteria library by tuning the threshold of the mapping score and maintaining a blacklist of concept clustering for different uses.

Other work closely related to DQueST lies in the research area of adaptive electronic questionnaire design. Though both aim to reduce the number of questions, most adaptive electronic (dynamic) questionnaires use association rules to predict part of the answers and, hence, shorten the questionnaire.[32,33] For example, if the answer to the question "male or female" is "male," then there is no need to ask a pregnancy-related question. On the other hand, DQueST, with the ultimate goal of clinical trial search optimization, reduces the number of questions by excluding noninformative questions, which is guided by the remaining trials in the working library. There are no conflicts to implement a parallel association rule-based question reduction module in DQueST to further shorten the questionnaire, which, however, is not the primary focus in this study.

Error analysis

We investigated the errors listed in Supplementary Material File S2, and summarized the errors made by DQueST, which fall into 3 categories. First, the complexities of unrestricted, hierarchical syntax and rich semantics in free-text eligibility criteria section in *ClinicalTrials.gov* makes it difficult to handle complex logic. As shown before, NegEx-based DQueST will mark "must not be pregnant" as "negated" in the exclusion criteria and will thus erroneously exclude patients who are not pregnant. In addition, DQueST treated all criteria in the inclusion section as "*required*" (ie, "AND" logic) which could cause potential errors. For example,

"Participants who have not received any systemic therapy for metastatic disease are also eligible" is not a corequirement of the preceding inclusion criteria ("AND" logic); instead it is an alternative criteria context that does not affect the preceding eligibility criteria context ("OR" logic).

Second, the high diversity of modifiers in eligibility criteria is another major barrier in generating questions. For example, an exclusion criterion could be "*Hypersensitivity to donepezil or piperidine.*" Three entities "*Hypersensitivity*," "*donepezil*," and "*piperidine*" will be recognized and 3 independent questions ignoring the relationship between them will be asked. Without understanding the modifier relationship between these entities, it is likely to raise errors in trial filtering.

Finally, DQueST assumes all criteria apply to the patient, which can cause errors since some criteria do not apply directly to the patients. For example, "A family member-caregiver (the medical decision-maker) for someone with Alzheimer's Disease" (eg, NCT00878059) is a criterion for recruiting caregivers instead of patients.

Limitations and future work

This study has a few limitations. One of the major deficits of our criteria library is its lack of subjective criteria. Subjective criteria, such as "*Only subjects who have participated in the previous study are eligible*," "*Capable of giving written informed consent*," and "*Subject able to be randomized within 18 hours of presentation*" (from NCT00297414), are commonly seen but are not stored in our criteria library since our named entity recognition model was trained with only clinically related terms. In order to incorporate nonclinical information, a more comprehensive annotated training corpus is required and the criteria representation model should be updated accordingly. In the future, we will generate a more comprehensive training data set with annotated relationships and use it to produce a cleaner and more comprehensive criteria library. We also would like to emphasize that the current web application is for demonstration purposes only, since this is a proof-of-concept study. An optimized production version is desired to increase the robustness and improve the user experience.

Another limitation of this study is that DQueST is confined to the eligibility criteria submitted to *ClinicalTrials.gov*. It is well known that those criteria are highly curated and often consist of only a subset of the entire criteria for a study. We observed that 20.6% trials contained no or less than 5 "parsable" criteria, a significant issue given DQueST is designed to filter trials by applying user answers to these criteria. On the other hand, while DQueST should theoretically be able to perform better with more comprehensive criteria, its accuracy depends on the NLP module in handling logic and entity recognition. Unfortunately, we were unable to compare DQueST's performance using the *ClinicalTrials.gov* criteria against internal protocol criteria as most of the clinical trial protocol documents are private and not easily obtained. However, we would like to emphasize that the 2 modules in DQueST work independently. Therefore, the criteria library (Module 1) can be manually curated and corrected in case high accuracy is desired, while the "question and answer" paradigm (Module 2) can still be used to filter out trials efficiently.

A third limitation of this study lies in its real-world evaluation. The number of patients recruited for conducting the usability evaluation is relatively small and large-scale tests are warranted to further improve the DQueST system. Additionally, the current design of

DQueST generates questions based on concepts defined in OMOP CDM, occasionally leading to confusion for patients without a medical background. Although patients can work alongside their clinicians to complete the questionnaire, our team aims to implement an info button specific to each medical concept to allow nonmedical users to complete this questionnaire without assistance.

CONCLUSION

DQueST uses dynamically generated questionnaires to guide clinical trial searches. It is an open-source system. It contributes a novel framework for transforming free-text eligibility criteria to questions and for filtering out trials dynamically based on user responses. It complements existing clinical trial search engines by bridging the gap between the limitations of keyword-based searches and the rich semantics embedded in free-text eligibility criteria. Future work needs to focus on identifying common data elements in eligibility criteria and curating a better library of standardized criteria representations.

FUNDING

AUTHOR CONTRIBUTIONS

CL, CY, and CW conceived the system design together. CL designed and implemented the system. CW supervised CL in design and implementation. CW and CT made critical revisions to the manuscript. CL, ZL, and AB contributed to the evaluation of the system. AB and RC contributed to the usability evaluation. CY contributed to the individual functions of the offline module. All authors edited and approved the manuscript.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

LICENSE

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Thadani SR, Weng C, Bigger JT, *et al*. Electronic screening improves efficiency in clinical trial recruitment. *J Am Med Inform Assoc* 2009; 16 (6): 869–73.
2. Biankin AV, Piantadosi S, Hollingsworth SJ. Patient-centric trials for therapeutic development in precision oncology. *Nature* 2015; 526 (7573): 361–70.
3. Zarin DA, Tse T, Williams RJ, *et al*. The ClinicalTrials.gov results database–update and key issues. *N Engl J Med* 2011; 364 (9): 852–60.
4. Miotto R, Jiang S, Weng C. eTACTS: a method for dynamically filtering clinical trial search results. *J Biomed Inform* 2013; 46 (6): 1060–7.
5. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004; 32: D267–70.
6. Weng C, Tu SW, Sim I, *et al*. Formal representation of eligibility criteria: a literature review. *J Biomed Inform* 2010; 43 (3): 451–67.
7. Tu SW, Peleg M, Carini S, *et al*. A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform* 2011; 44 (2): 239–50.
8. Luo Z, Miotto R, Weng C. A human-computer collaborative approach to identifying common data elements in clinical trial eligibility criteria. *J Biomed Inform* 2013; 46 (1): 33–9.
9. Weng C, Wu X, Luo Z, *et al*. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc* 2011; 18 (Suppl 1): i116–24.
10. Boland MR, *et al*. EliXR-TIME: a temporal knowledge representation for clinical research eligibility criteria. *AMIA Jt Summits Transl Sci Proc* 2012; 2012: 71–80.
11. Hao T, Liu H, Weng C. Valx: a system for extracting and structuring numeric lab test comparison statements from text. *Methods Inf Med* 2016; 55 (3): 266–75.
12. 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.
13. Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition. Journal of the American Medical Informatics Association. 2019 Feb 7;26(4):294–305.
14. Obeid JS, McGraw CA, Minor BL, *et al*. Procurement of shared data instruments for research electronic data capture (REDCap). *J Biomed Inform* 2013; 46 (2): 259–65.
15. Rothman KJ, Mikkelsen EM, Riis A, *et al*. Randomized trial of questionnaire length. *Epidemiology* 2009; 20 (1): 154.
16. Edwards P. Questionnaires in clinical trials: guidelines for optimal design and administration. *Trials* 2010; 11 (1): 2.
17. Manning C, *et al*. The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. June 23-24, 2014 Baltimore, Maryland, USA.
18. Lafferty J, McCallum A, Pereira FC. *Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data*. In: *Proceedings of the 18th International Conference on Machine Learning 2001 (ICML 2001)*, pages 282–289. http://portal.acm.org/citation.cfm?id=655813
19. Chapman WW, Bridewell W, Hanbury P, *et al*. A simple algorithm for identifying negated findings and diseases in discharge summaries. *J Biomed Inform* 2001; 34 (5): 301–10.
20. Hripcsak G, *et al*. Observational health data sciences and informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015; 216: 574–8.
21. Si Y, Weng C. An OMOP CDM-based relational database of clinical research eligibility criteria. *Stud Health Technol Inform* 2017; 245: 950–4.
22. Białecki A, Muir R, Ingersoll G. Apache Lucene 4. In: SIGIR 2012 Workshop on Open Source Information Retrieval; 2012.
23. Brooke J. SUS: a retrospective. *J Usability Stud* 2013; 8 (2): 29–40.
24. Simmons MF, Laney DA. ThinkGenetic: identification of misinformation and educational gaps using an innovative and interactive website. *Mol Genet Metab* 2017; 120 (1-2): S123.

25. Yun BH, Seo CH. Semantic-based information retrieval for content management and security. *Comput Intell* 2003; 19 (2): 87–110.

26. Bhupatiraju RT, Fung KW, Bodenreider O. MetaMap lite in excel: biomedical named-entity recognition for non-technical users. *Stud Health Technol Inform* 2017; 245: 1252.

27. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010; 17(3):229–36.

28. Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.

29. Friedman C, Alderson PO, Austin JH, *et al.* A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994; 1 (2): 161–74.

30. Collobert R, Weston J. A unified architecture for natural language processing: deep neural networks with multitask learning. In: Proceedings of the 25th International Conference on Machine Learning. Helsinki, Finland. July 05 - 09, 2008.

31. Milian K, Hoekstra R, Bucur A, *et al.* Enhancing reuse of structured eligibility criteria and supporting their relaxation. *J Biomed Inform* 2015; 56: 205–19.

32. Ortigosa A, Paredes P, Rodriguez P. AH-questionnaire: An adaptive hierarchical questionnaire for learning styles. *Comput Educ* 2010; 54 (4): 999–1005.

33. Mwamikazi E, Fournier-Viger P, Moghrabi C, An adaptive questionnaire for automatic identification of learning styles. In: International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. Kaohsiung, Taiwan, June 3-6, 2014.