A Framework for Systematic Assessment of Clinical Trial Population Representativeness Using Electronic Health Records Data

Yingcheng Sun[1] Alex Butler[1,2] Ibrahim Diallo[1] Jae Hyun Kim[1] Casey Ta[1] James R. Rogers[1] Hao Liu[1] Chunhua Weng[1]

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

Address for correspondence Chunhua Weng, PhD, Department of Biomedical Informatics, Columbia University, 622 West 168 Street, PH-20 room 407, New York, NY 10032, United States (e-mail: chunhua@columbia.edu).

Appl Clin Inform 2021;12:816–825.

Abstract

Background Clinical trials are the gold standard for generating robust medical evidence, but clinical trial results often raise generalizability concerns, which can be attributed to the lack of population representativeness. The electronic health records (EHRs) data are useful for estimating the population representativeness of clinical trial study population.

Objectives This research aims to estimate the population representativeness of clinical trials systematically using EHR data during the early design stage.

Methods We present an end-to-end analytical framework for transforming free-text clinical trial eligibility criteria into executable database queries conformant with the Observational Medical Outcomes Partnership Common Data Model and for systematically quantifying the population representativeness for each clinical trial.

Results We calculated the population representativeness of 782 novel coronavirus disease 2019 (COVID-19) trials and 3,827 type 2 diabetes mellitus (T2DM) trials in the United States respectively using this framework. With the use of overly restrictive eligibility criteria, 85.7% of the COVID-19 trials and 30.1% of T2DM trials had poor population representativeness.

Conclusion This research demonstrates the potential of using the EHR data to assess the clinical trials population representativeness, providing data-driven metrics to inform the selection and optimization of eligibility criteria.

Keywords
- clinical trials
- eligibility criteria
- generalizability assessment
- population representativeness
- information extraction
- natural language processing

Background and Significance

Clinical trials generate evidence about the effectiveness, efficacy, or safety of new treatments.[1] The success of a trial hinges on timely recruitment of adequate representative patients.[2,3] Overly restrictive eligibility criteria can limit the representativeness of study samples, and lead to low participation that can delay the trial, lead to its termination, or cause safety issues.[4] Trial designers often rely on previous clinical trials and past experience for patient selection, but this type of selection process can be subjective and lack transparent rationale.[5,6] Van Spall et al[7] reviewed 283 clinical trials and found that 84.1% of them contained at least one poorly justified exclusion criterion. Whether these experiments can be extrapolated to broader populations is uncertain, and the compromised generalizability of clinical studies is a long-standing concern.

Methods have been developed to quantify population representativeness. *A posteriori* generalizability method was developed to retrospectively examine clinical trial

received
April 18, 2021
accepted after revision
June 23, 2021

Georg Thieme Verlag KG,
Rüdigerstraße 14,
70469 Stuttgart, Germany

DOI https://doi.org/10.1055/s-0041-1733846.
ISSN 1869-0327.

population representativeness after study completion. Janson et al[8] examined the representativeness of a colon cancer laparoscopic or open resection trial by comparing included and excluded patients in the participating Swedish centers. Van der Aalst et al evaluated the degree of self-selection in a Dutch–Belgian randomized controlled lung cancer screening trial to assess the generalizability of the study results.[9] Bress et al studied the generalizability of the Systolic Blood Pressure Intervention Trial (SPRINT) in detail using data from the National Health and Nutrition Examination Survey and found a substantial percentage of U.S. adults met the eligibility criteria for SPRINT.[10] Such *a posteriori* generalizability analyses are unable to provide early intervention during trial design.

In contrast, the *a priori* generalizability assessment is an eligibility criteria-driven analysis conducted before the trial commences and can potentially provide early estimation of the population representativeness to enable iterative refinement of trial eligibility criteria. Weng et al proposed a quantitative metric Generalizability Index for Study Traits (GIST) to quantify the proportion of patients that would be potentially eligible across trials with the same clinical trait over the target population.[11] This method can correlate adverse events with criteria's population representativeness.[12] Later, Sen et al[13] extended GIST to GIST 2.0 as a quantitative metric to assess the *a priori* generalizability based on population representativeness of a clinical trial by accounting for the dependencies among multiple eligibility criteria. Cahan et al proposed a metric that computes the similarity between the study population and the target population one characteristic a time.[14] Despite the methodological advances contributed by the above studies, there has been no integrated analytical pipeline that enables end-to-end automatically systematic analysis of clinical trial population representativeness.

Objectives

In this paper, we contribute an automated analytical framework that leverages natural language processing (NLP) technologies to quantify clinical trial population representativeness. Discrete eligibility traits are extracted from free-text eligibility criteria, and each trait represents a single eligibility rule conforming to the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM).[15] Queries are then formulated based on the eligibility rules and executed on standardized clinical data to construct a study cohort. Quantitative metrics accounting for the difference of patients are provided based on nonlinear regression. We applied the framework to 782 novel coronavirus disease 2019 (COVID-19) trials and 3,827 type 2 diabetes mellitus (T2DM) trials to evaluate the population representativeness, respectively.

Methods

The clinical trial protocol includes detailed experiment introduction and requirements, which can be classified into metadata and eligibility criteria. Metadata defines indexing characteristics of clinical trials, such as study type, intervention, medical condition, and study design. Eligibility criteria define rules specifying who is eligible (inclusion criteria) or ineligible (exclusion criteria) to participate in a clinical trial and are usually documented as free-text format.[16] ►Fig. 1 provides an overview of the analytical framework including eligibility criteria annotation and normalization, query formulation, and representativeness assessment modules.

Criteria Annotation and Normalization

To transform the eligibility criteria from free-text into a structured and normalized format, we performed Named Entity Recognition (NER) to extract and normalize concepts in criteria text using an open-source tool called Critera2Query.[17] Compared with other NER tools such as cTAKES[18] or MetaMap,[19] Criteria2Query maps concepts into the OMOP CDM with considerable precision (~90%) and recall (~71%). The OMOP CDM is an open-source community standard for observational healthcare data, with stronger international orientation than any other data model.[15] It receives regular terminology updates and is widely used in the global scientific community for health data standardization.[20,21] Entities in the criteria are automatically recognized and classified into one of the six domains:

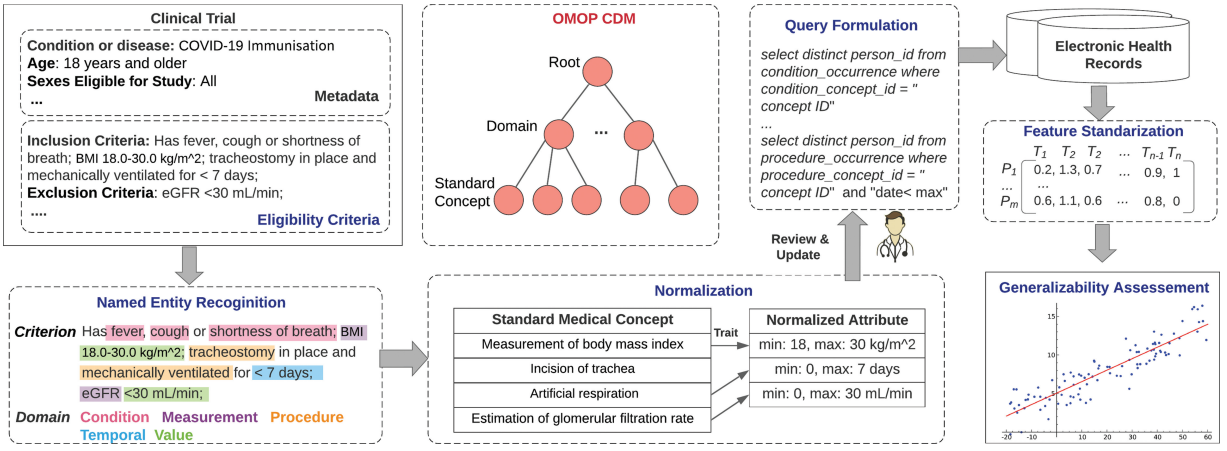

Fig. 1 Overview of the analytical framework. It takes free-text eligibility criteria as input, goes through Named Entity Recognition, Concept Normalization and Query Formulation, all using the OMOP CDM, and executes the cohort query using EHR data and reports representativeness of the study cohort in the EHR population visually. EHR, electronic health record.

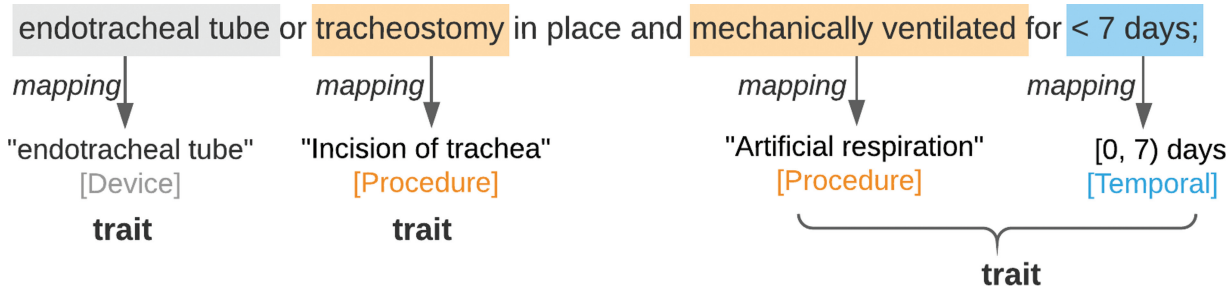

Fig. 2 Example traits, which can be an entity or an entity with its attribute, in a criterion.

observation, drug, person, procedure, measurement and condition, and then mapped to standard medical concepts in the OMOP vocabulary. Entities can be described by their attributes, which define a range of values that the entities might hold. A value or temporal type of attribute is usually used in the eligibility criteria and normalized together with corresponding entities. The value attributes are normalized as numerical values with upper and lower boundaries. For example, "BMI of 18.0 to 30.0 kg/m^2" is coded with a minimum boundary as "18.0" and maximum boundary as "30.0" for the entity "measurement of body mass index (BMI)." Regular expressions are used to identify and convert different type of operators. The temporal attributes are unified to the same unit (days) by SUTime from standard NLP group.[22] For example, "for at least 1 year before the screening visit" is coded into "≥365 days before the screening visit." After normalizations, the attributes are converted from strings to numerical data types and can be comparable in a quantitative manner.

Each eligibility criterion contains one or more traits. A criterion with three traits is shown in ►Fig. 2. A normalized entity and its attribute form a "trait," like "artificial respiration within (0–7) days." If there is no attribute, the normalized entity itself can be a trait, such as the normalized entity "endotracheal tube" or "incision of trachea."

Query Formulation

Queries are automatically formulated with traits to identify eligible cohorts from an OMOP CDM clinical database. ►Fig. 3 shows the optimized query formulation and cohort discovery process.

First, all patients eligible for a study condition or disease (usually declared in the metadata of clinical trials) will be selected from the EHR database to create the initial target population. If a patient has multiple instances of record, only the latest one will be counted so that a patient will enter a cohort only once. Distinct eligibility traits are used to formulate queries. Temporary tables with eligible patients for each distinct trait under the target population are created to reduce the cost of repeatedly querying the original tables when the same trait is defined in different criterion. The overall eligible patients are the union of eligible patients for traits connected by "OR" relationship, and intersection of eligible patients for other traits. The study cohort of a clinical trial is the intersection of eligible patients for all included eligibility criteria. Exclusion criteria will be automatically negated in the query formulation. The "join" clause is used to combine multiple tables in the database.

Representativeness Assessment

The goal of the representativeness assessment is to determine which traits have the strongest effect on the overall representativeness of a trial. A few existing models[11,23] measure the population representativeness by simply calculating the fraction of the number of eligible patients, but that could fail to distinguish patients in terms of their trait values. For example, assume we have two traits specified in a clinical trial and each

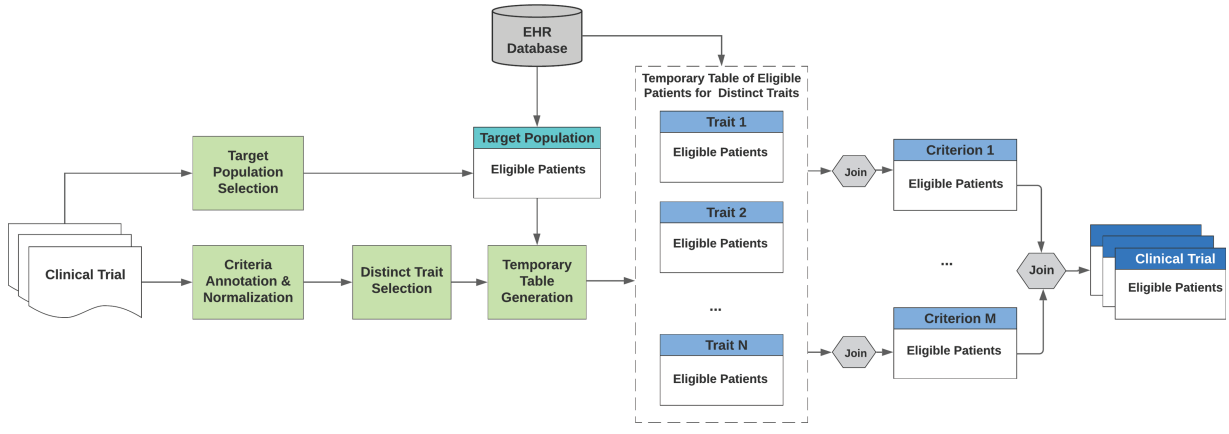

Fig. 3 The query formulation and cohort identification process. The framework supports cohort identification for multiple trials with the same study condition in a batch mode.

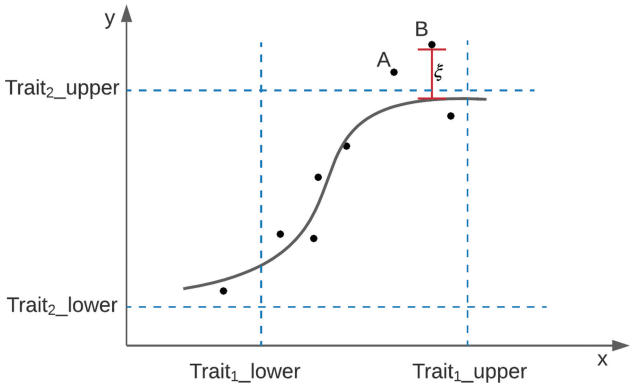

Fig. 4 Example of a nonlinear regression model with two traits. $Trait_i$_lower and $Trait_i$_upper represent the lower and upper bounds of trait *i*. Each dot is a patient represented by a two-dimensional vector.

of them has attribute with lower and upper bounds as ►Fig. 4 shows. Each patient can then be represented by a two-dimensional vector and each trait is a feature vector. Suppose patient A and B are eligible for trait 1 and ineligible for trait 2, but their recorded values for trait 2 are different. If we move the upper bound of trait 2 higher to cover more patients, patient A will need less boundary relaxation than patient B, but only counting the number of eligible patients cannot model such difference.

EHR data contain large amounts of information about patients' medical history including symptoms, examination findings, test results, prescriptions, and procedures that often have a nonlinear association.[24] Support Vector Regression (SVR)[25] model is an effective method to learn a nonlinear model with a hyperplane.[26] First, the trait feature vectors are pre-possessed by removing the mean and scaling to unit variance for standardization. Next, a robust support SVR model with radial basis function (RBF) kernel is learned from the standardized features, and the weight of each patient is calculated by Eqs. 123.

Equations

$$min\ \frac{1}{2}\ ||w||^2 \tag{1}$$

subject to

$$|y_i - w^T x_i| \leq \text{å}, i = 1 \ldots N \tag{2}$$

$$\ddot{a}_i = \frac{1}{1 + |y_i - w^T x_i|} \tag{3}$$

For patient *i* with multiple traits, one trait is designated as the dependent variable y_i, and all others are designated as independent variables x_i to compute a hyperplane, *w* is the is the normal vector to the hyperplane, and å defines a margin of tolerance with default value as 0, $\ddot{a}_i$ is the weight assigned to patient *i* that is inversely proportional to the residual distance from the patient data point *i* to the hyperplane. We define *g_t* and *g_l* as the metrics for the population representativeness of a trait and trial in Eqs. 4 and 5.

$$g_{t_k} = \frac{\Sigma_{patient\ j \in EP_{trait_k}} \ddot{a}_j}{\Sigma_{patient\ l \in EP_{pool}} \ddot{a}_l} \tag{4}$$

$$g\_l = \frac{\Sigma_{patient\ i \in EP\_trial} \ddot{a}_i}{\Sigma_{patient\ l \in EP\_pool} \ddot{a}_l} \tag{5}$$

where EP_trait_k is the set of eligible patients of trait *k*, *EP_pool* is the union of eligible patients of each trait and *EP_trial* is the intersection of eligible patients of each trait. *EP_trial* represents the target cohort of the clinical trial and *g_l* measures its population representativeness. The *g_t* or *g_l* score is always between 0 and 1, with higher score implying greater population representativeness. A score of 0 means no patient was found eligible for the selected trait or trial. In contrast, a score of 1 for g_t_k means all patients with trait *k* were eligible for the clinical trial, and a score of 1 for *g_l* means the eligible patients for each trait are the same as the eligibility criteria of the trial.

By training a regression model in analytical framework, we can obtain the hyperplane and the average distance from all points to this hyperplane is the shortest. In Eqs. 3, outliers receive less weights, so the patients whose trait observations differ significantly from others may influence the population representativeness more than patients close to the hyperplane. If the *g_l* score is low and there are many outliers against the hyperplane for a trait, the relaxation of the corresponded criterion could improve the population representativeness, so *g_l* score is a trait attention-guided metric for the general clinical trial population representativeness assessment.

Results

We applied the analytical framework on clinical trials of two different diseases and assessed the population representativeness for each trial by calculating its *g_l* score and identifying traits that limit the trial population representativeness through *g_t* scores.

Datasets

We chose the clinical studies on two different diseases as our datasets COVID-19 and T2DM. Since the first reported case in December 2019, COVID-19 has spread rapidly from country to country and become one of the worst pandemics in the world's history.[27,28] T2DM, recognized as an important public health problem by the World Health Organization, can lead to serious damage to the heart, blood vessels, eyes, kidneys, and nerves.[29] A total of 782 interventional COVID-19 and 38,27 T2DM clinical trials with recruiting site in the United States were exported from ClinicalTrials.gov in July 2020 by querying the study condition "COVID-19" and "type 2 diabetes mellitus," respectively. The study population data are from Columbia University Irving Medical Center EHR data that have been converted and stored in the OMOP CDM

Table 1 Total count and percentage of entities extracted from inclusion and exclusion criteria in COVID-19 and T2DM clinical trials

Criteria type	COVID-19		T2DM	
	Count	%	Count	%
Inclusion	6,068	35.79%	23,194	40.29%
Exclusion	10,909	64.21%	34,377	59.71%
Total	16,977	100.00%	57,571	100.00%

Abbreviations: COVID-19, novel coronavirus disease 2019; T2DM, type 2 diabetes mellitus.

Table 2 Total count and percentage of entities in different domains for COVID-19 and T2DM clinical trials

Domain	COVID-19		T2DM	
	Count	%	Count	%
Condition	6,601	38.88%	28,138	48.88%
Observation	3,946	23.24%	769	1.34%
Drug	2,199	12.95%	9,770	16.97%
Measurement	2,164	12.75%	10,681	18.55%
Procedure	1,372	8.08%	559	0.97%
Person	695	4.09%	7,654	13.29%
Total	16,977	100.00%	57,571	100.00%

Abbreviations: COVID-19, novel coronavirus disease 2019. T2DM, type 2 diabetes mellitus.

format by querying the condition occurrence in "COVID-19 (ID 37311061)" and "type 2 diabetes mellitus (ID 201826)." We identified 9,664 COVID-19 patients and 54,273 T2DM patients.

Eligibility Criteria Normalization

The free-text eligibility criteria in the exported clinical trials were translated to structured representations through the entity recognition and normalization process. Each entity was labeled as one of the following domains: procedure, measurement, drug, condition, observation, and person. The "person" domain includes two demographic criteria "age" and "sex" for each trial. ►Tables 1 and 2 list the statistical information for the extracted entities. Every extracted entity was mapped to a standard medical concept, and, if applicable for the entity, its attributes were normalized. Different entities with the same meaning may be mapped to the same concept.[30] Three common types of concept mappings (exact, partial, and semantic) were presented in the ►Table 3.

Systematic Trial Population Representativeness Assessment

With the normalized eligibility criteria, queries were automatically formulated and then we applied the *g_l* metric to assess the population representativeness for all COVID-19 and T2DM clinical trials. ►Figs. 5 and 6 show the distributions of *g_l* scores for COVID-19 and T2DM trials. The average and median *g_l* scores for all COVID-19 trials are 0.065 and 0.02, respectively. Most (85.7%) trials' *g_l* scores locate on the low score area (less than 0.1), and very few trials have high population representativeness. The average and median *g_l* scores for all T2DM trials is 0.1 and 0.03, respectively. The *g_l* scores of 29.3% trials equal to zero and 0% trials are one, which is roughly consistent with the experimental results produced by Sen et al.[31] To improve the trial population representativeness, the analytical framework provides the option to drop the traits with zero *g_t* score. We removed the traits with zero *g_l* score from the eligibility criteria and recalculate the *g_l* score for every trial and assess the population representativeness again. From the ►Fig. 5, we can see there is an obvious increasement (the red color) of the *g_l* scores, that means the population representativeness of trials are improved. There are still 50.9% trials less than 0.1 that is because there are no enough COVID patients yet in the database as of the time for this study. As more and more EHRs imported, there will be more eligible patients for various traits in the database. From the ►Fig. 6, we can see there is an obvious increasement (the red color) of the *g_l* scores, meaning the population representativeness of T2DM trials are improved.

Table 3 Examples of extracted entities with their mapped concept and normalized attribute.

Entity	Attribute	Domain	Concept ID	Concept name	Min. value	Max. value
Shortness of breath	None	Condition	312437	Dyspnea		
History of	None	Observation	4188893	History of		
Systemic corticosteroids	exceeding 10 mg/day of prednisone equivalent	Drug	21605200	Corticosteroids	10	Infinite
Weight	>3 and <40 kg	Measurement	3025315	Body weight	3	40
Mechanical ventilation	>24 h	Procedure	4230167	Artificial respiration	24	Infinite
Age	18–65 y	Person	4156190	Age	18	65

Note: As presented in the table, the mapping procedure can vary based on the starting entity (e.g., "history of" and "age" are exact mappings; "systemic corticosteroids" and "weight" are partial mappings; and "shortness of breath" and "mechanical ventilation" are completely semantic mappings).

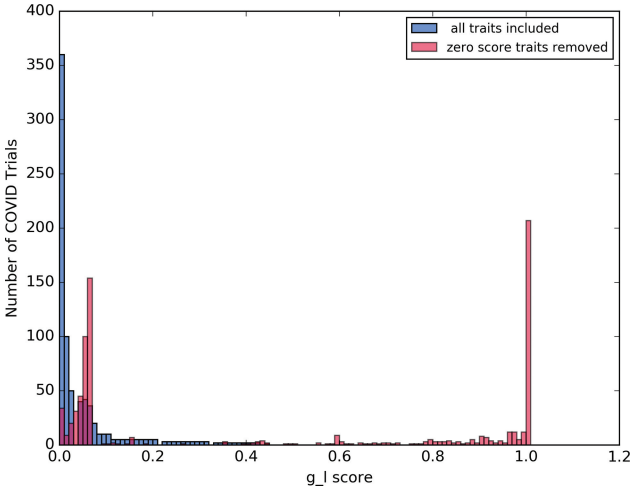

Fig. 5 Histogram of the *g_l* score values for COVID-19 trials. The blue color represents the trials with all traits included and the red color represents the "relaxed" trials with zero *g_t* score traits removed.

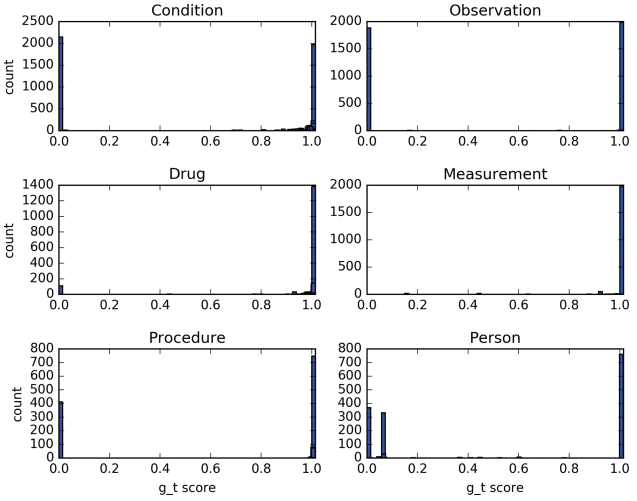

Fig. 7 Histogram of *g_t* scores for traits in six domains for COVID-19 clinical trials. The y-axis indicates the count of traits.

To figure out which types of traits limit the trial population representativeness, we calculate the *g_t* score for each single trait and aggregate them by domains. ►**Figs. 7** and **8** show the histogram of the *g_t* scores for all 16,977 traits for COVID-19 clinical trials and 57,571 traits for T2DM clinical trials in various domains. For COVID-19 trials, 32.48% traits in condition domain, 47.74% in observation domain, 4.87% in drug domain, and 29.96% in procedure domain have *g_t* scores less than 0.1. For T2DM clinical trials, 16.7% traits in condition domain, 31.7% traits in observation domain, 28.5% traits in drug domain, 25.4% traits in procedure domain, and 17.1% traits in person domain have *g_t* scores less than 0.1.

The population representativeness of a trial depends on the union of all traits it contains, so the traits with low *g_t* scores will be directly related to the decrease of the overall population representativeness, especially the traits with a *g_t* score equal to 0. For the traits consisted of single medical concept, the low *g_t* score is caused by no enough events occurred in the EHR database. For example, patients who died in the nursing home or long-term care facility after discharge from the hospital cannot be identified in current OMOP CDM, so the *g_t* scores for the observation "death in nursing home or long-term care facility" will be 0. For the traits composed of a medical concept together with its attribute, the low *g_t* scores could be caused by either no enough existence of records related to that specific concept or too few records eligible for the condition that the attribute specifies. For example, almost all patients' "age" information were recorded in the database, but very few people "older than 65." There are many "HbA1c" tests ordered and the results were imported to the database, but results with "HbA1c less than 7" are few, which is the value range specified in the eligibility criteria of several T2DM clinical trials.

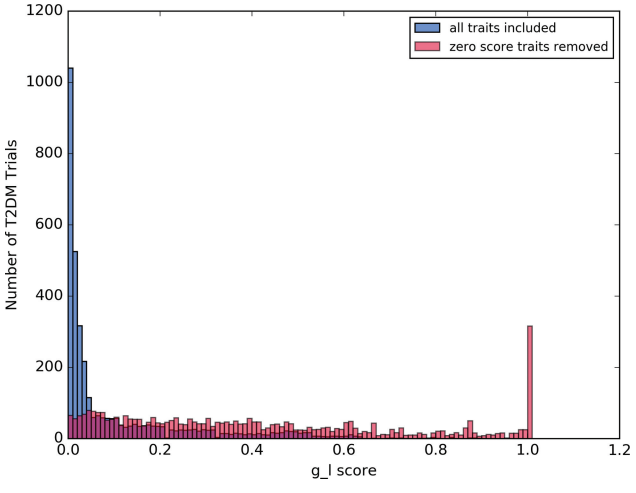

Fig. 6 Histogram of the *g_l* score values for T2DM trials. The blue color represents the trials with all traits included and the red color represents the "relaxed" trials with zero *g_t* score traits removed.

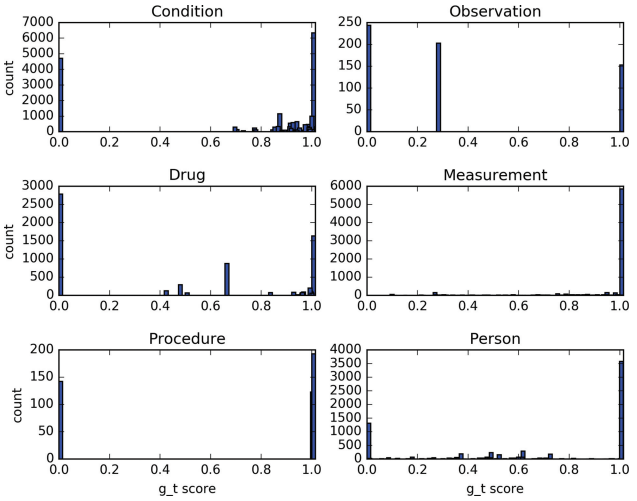

Fig. 8 Histogram of *g_t* scores for traits in six domains for T2DM clinical trials. The y-axis indicates the count of traits.

Table 4 Top 5 criteria with highest and lowest patient representativeness in COVID-19 trials

	Criteria	INC/EXC	*g_t* score	No. of eligible patients
Criteria with highest patient representativeness	*Detection of 2019 novel coronavirus using polymerase chain reaction technique* (37310255)	INC	1.0	9,664
	Ages (4156190) eligible for study: 12 y and older	INC	1.0	9,664
	Sexes (4135376) eligible for study: all	INC	1.0	9,664
	Fever (437663)	INC	1.0	9,664
	Cough (254761)	INC	1.0	9,664
Criteria with lowest patient representativeness	*Patient currently pregnant* (4299535)	INC	0.002	<100
	History of active or treated *lung cancer* (255573)	INC	0.001	<50
	Patient receiving angiotensin converting enzyme (ace) inhibitor or angiotensin receptor blocker (arb) therapy (2617905)	INC	0	<10
	Coronary artery bypass graft (4336464)	INC	0	<10
	Prediabetes (37018196)	INC	0.001	<10

Abbreviations: COVID-19, novel coronavirus disease 2019. INC, inclusion. EXC, exclusion.
Note: Entities are in italic and followed by standard concept IDs.

Evaluation of Population Representation Assessment

To evaluate the analytical framework, we first measure the correlation between patient representativeness metric for the signal trait (*g_t* score) and the number of eligible patients, and the results show they are positively correlated ($r = 0.8103, p < 0.0001$). ►Tables 4 and 5 list the top 5 criteria with highest and lowest patient representativeness in COVID-19 and T2DM trials, respectively. Next, we evaluate the *g_l* score on population representation assessment and relaxation. Considering the labor cost, we randomly picked two clinical trials from the COVID-19 and T2DM trial corpus respectively for evaluation. ►Tables 6 and 7 present the representation assessment results and relaxation process.

After initial patient representation assessment for the COVID-19 trial (NCT04540406), the *g_l* score was 0.001 meaning the population representativeness is very poor, and only a few patients were eligible for the trial (#*p_l* < 20). Among all the traits, "have prediabetes or T2DM" had the lowest population representativeness (*g_t* = 0.002) that lowered the overall trial population representativeness. This might be caused by the data sparsity with no enough COVID-19 patients with "prediabetes or T2DM" had been recorded in the database at the time of experiment. We manually removed this trait in the first round of relaxation, and then the *g_l* score was increased to 0.82 with acceptable population representativeness and 9,013 patients were found eligible. On the contrary, if we removed a trait with high *g_t* score like "no extracorporeal membrane oxygenation" for the trial, the population representativeness will not change because it is derived from the intersection of all traits and depends on the most restrictive one.

Table 5 Top 5 criteria with highest and lowest patient representativeness in T2DM trials

	Criteria	INC /EXC	*g_t* score	No. of eligible patients
Criteria with highest patient representativeness	Diagnosed with *type 2 diabetes* (380096)	INC	1.0	54,273
	Gender (4135376): all	INC	1.0	54,273
	HbA1c (2212392) measurement	INC	1.0	54,273
	Drug abuse (436954)	EXC	0.99	54,262
	Ages (4156190) Eligible for Study: 18 Years and older	INC	0.99	54119
Criteria with lowest patient representativeness	unstable *chronic disease* (443783)	EXC	0.0003	<100
	oral antidiabetic drug: *α-glucosidase inhibitors* (1169352)	INC	0.001	<50
	Age (4156190):9–16	INC	0.002	<50
	Female of childbearing age (4142985)	INC	0.001	<10
	using adequate *contraception* (4027509)	INC	0	<10

Abbreviations: T2DM, type 2 diabetes mellitus; INC, inclusion; EXC, exclusion.
Note: Entities are in italic and followed by standard concept IDs. To protect patient privacy, all criteria where the count of eligible patients less than 100 were not provided with the actual number.

Table 6 List of population representation assessment results for COVID-19 trial NCT04540406

Trait	Initial Assessment *g_l* = 0.001, #p_l < 20		1st relaxation: *remove "have prediabetes or T2DM"* *g_l* = 0.82, *#p_l* = 9,013		2nd relaxation: *remove "No Extracorporeal membrane oxygenation"* *g_l* =0.82, *#p_l* =9,013	
	g_t	*#p_t*	*g_t*	*#p_t*	*g_t*	*#p_t*
Age >= 18	0.94	9,057	0.94	9,057	0.94	9,057
Sex: all	1.00	9,664	1.00	9,664	1.00	9,664
Disease caused by severe acute respiratory syndrome coronavirus 2	1.00	9,664	1.00	9,664	1.00	9,664
No Blood transfusion reaction	0.99	9,654	0.99	9,654	0.99	9,654
No Extracorporeal membrane oxygenation	0.98	9,621	0.98	9,621		
Have prediabetes or T2DM	0.002	< 20				

Abbreviations: COVID-19, novel coronavirus disease 2019. T2DM, type 2 diabetes mellitus.
Note: *#p_t*: the number of eligible patients for a specific trait, *#p_l*: the number of eligible patients for all traits. "Relaxed" traits are in italics. To protect patient privacy, all criteria where the count of eligible patients less than 100 were not provided with the actual number.

Table 7 List of population representation assessment results for T2DM clinical trial NCT01032629

Trait	Initial assessment *g_l* = 0.04, *#p_l* = 854		1st relaxation: *remove "cardiovascular disease"* *g_l* =0.22, *#p_l* =10,457		2nd relaxation: *update HbA1c: (5.7–10)* *g_l* =0.31, *#p_l* =*22,243*	
	g_t	*#p_t*	*g_t*	*# p_t*	*g_t*	*#p_t*
Age >50	0.94	44,689	0.94	44,689	0.94	44,689
Sex: all	1.00	54,273	1.00	54,273	1.00	54,273
No sitagliptin	0.99	54,244	0.99	54,244	0.99	54,244
No Ketoacidosis	1.00	54,273	1.00	54,273	1.00	54,273
No Type 1 diabetes mellitus	0.95	48,291	0.95	48,291	0.95	48,291
HbA1c: (6.5–8)	0.34	13,162	0.34	13,162	0.42	28,534
Cardiovascular disease	0.09	3,781				

Abbreviations: T2DM, type 2 diabetes mellitus.
Note: *#p_t*: the number of eligible patients for a specific trait, *#p_l*: the number of eligible patients for all traits. "Relaxed" traits are in italics.

After the initial assessment for the T2DM clinical trial (NCT01032629), the population representativeness was poor (*g_l* = 0.04), and 854 patients were found eligible. The trait "cardiovascular disease" had the lowest *g_t* score, and we tried to manually remove it in the first round of relaxation, and then the trial had better population representativeness and the number of eligible patients was increased to 10,457. The trait "*HbA1c*" is usually used in T2DM trials as the primary defining trait, but sometimes its value range might be a little restrictive. Compared with other traits, "HbA1c" had relatively lower *g_t* score, so its range was relaxed from (6.5–8) to (5.7–10) in the second round of relaxation, and then the number of eligible patients was almost doubled.

The target cohort needs to meet all the eligibility traits (rules) in the clinical trial, so overly restrictive traits will make the targe cohort less representative with fewer patients. The *g_t* and *g_l* scores can be useful metrics in qualifying the trial population representativeness assessment and indicators for identifying restrictive traits and guiding the loosening of eligibility criteria to improve the clinical trial population representativeness.

The analytical framework improves the computational efficiency by using temporary tables and processing trials in a batch mode. We evaluated the performance of the analytical framework by analyzing the computation time for calculating the population representativeness score and compared the results with that from using GIST 2.0 on an iMac computer with Intel Core I7 (4.2 Ghz) CPU, 16 GB 2400 MHz DDR4 memory, and 1TB SSD hard disk. On average, GIST 2.0 spends 1,332.85 seconds in total for population representativeness score calculation, while the proposed framework only needs 101.64 seconds, which is 93.4% faster than GIST 2.0.

Discussion

We have presented an end-to-end systematic population representative analytical framework and applied it to clinical

studies on real patient EHRs. Overly restrictive eligibility criteria were identified and filtered to improve the clinical trial population representativeness. The population representativeness assessment consists of multiple steps, and there are potential reasons that may affect representativeness assessment results.

Inaccurate or incomplete criteria extraction might limit the automatic population representativeness analysis by either underestimating or overestimating the population representativeness score. Due to the semantic complexity, it is hard for automatic information extraction tools to reach 100% accuracy in either NER or criteria normalization.[32,33] For traits with numeric values, the possible unit difference between the criteria and EHR data might also impede the population representativeness evaluate. In this study, these inaccurate annotations were manually updated by domain experts. The updated information extraction pipeline Criteria2Query (http://34.70.212.14:8080/criteria2query_box/) is enhanced with an editable user interface so that human experts can manually review and edit the output for better accuracy. Moreover, for all the trials analyzed in this study, two reviewers (A.B. and J.R.) examined the results manually and recommended updates as needed. A medical terminology search engine Athena (http://athena.ohdsi.org) that provides searching of concepts in the OMOP CDM is used to assist experts in spot check in case of doubt for some concepts. All processed criteria were reviewed by a third annotator (Y.S.) to check for and resolve discrepancies.

The ambiguity of eligibility criteria themselves also lead to the failure of the clinical trial population representativeness. For example, in the exclusion criteria "with clinical manifestation of renal impairment (e.g., a creatinine value of 1.5 times or more of the upper reference limit (NCT01318135)," the "creatinine value" is described as "upper limit of normal" without declaring the exact value by trial conductors. The normal range of some laboratory measurements are not unique and dependent on institutions and patient characteristics such as sex. Without explicit mention of the reference range, it might result in unreal estimation of trial population representativeness.

Different institutions may use different medical concept coding systems or data models, and the concepts extracted from the eligibility criteria by the analytical framework may not be successfully mapped to the patient EHR data from other systems. In our EHR data, all medical concepts have been mapped to the standard concepts in the OMOP CDM. For example, the blood test measurement "hemoglobin A1c" were coded as multiple concepts in the original EHRs such as "hemoglobin A1c/hemoglobin.total in blood (LOINC 3004410)," "hemoglobin A1c/hemoglobin.total in blood by HPLC (LOINC 3005673)" and "deprecated hemoglobin A1c in blood (LOINC 40758583)," but all converted to the standard concept "hemoglobin; glycosylated (A1C) (2212392)" in the OMOP CDM compliant version. It may fail in cohort discovery for "hemoglobin A1c" coded by other data models.

Eligibility criteria need to be clinically justified so that low representativeness scores (*g_m* or *g_l*) do not necessitate the need to loosen a criterion. This framework is designed to serve as a decision aid for clinical trial designers by improving the transparency of population representativeness of individual criteria and alerting clinical experts of potentially unrepresentative or restrictive criteria. The output of each module, including eligibility criteria normalization, query formulation, population representativeness assessment, is all preserved and formatted for convenient verification by domain experts.

Conclusion

In this paper, we contributed an integrated clinical trial population representativeness assessment pipeline and applied it on COVID-19 trials and T2DM trials for representativeness assessment. A few limitations that may compromise the representativeness estimation accuracy were discussed. As our future research, we plan to improve the medical entity recognition and hierarchical concept mapping for more accurate population representativeness assessment.

Clinical Relevance Statement

The success of a trial depends on the inclusion of adequate participants. Leveraging EHR data to support the clinical trial design process addresses the need of investigators to optimize the eligibility criteria to avoid poor population representativeness. Automatic and systematic eligibility criteria normalization and query formulation framework reduces the human labor when evaluating the population representativeness of clinical trials.

Multiple Choice Questions

1. What are queries formulated with to identify eligible cohorts from the clinical database?
   a. Entity
   b. Attribute
   c. Trait
   d. Domain

   Correct Answer: The correct answer is option c. Queries formulated with traits to identify eligible cohorts. A trait can be an entity or an entity with its attribute in a criterion.

2. What does a high *g_l* score mean?
   a. Great population representativeness of a clinical trial
   b. Poor population representativeness of a clinical trial
   c. Great population representativeness of an eligibility criterion
   d. Poor population representativeness of an eligibility criterion

   Correct Answer: The correct answer is option a. g_l is the metric to measure the population representativeness of a clinical trial with its value between 0 and 1, with higher scores representing greater representativeness.

Protection of Human and Animal Subjects

No human or animal subjects were involved in the project.

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.

Conflict of Interest

None declared.

References

1 Piantadosi S. Clinical Trials: A Methodologic Perspective. John Wiley & Sons; 2017
2 Fogel DB. Factors associated with clinical trials that fail and opportunities for improving the likelihood of success: a review. Contemp Clin Trials Commun 2018;11:156–164
3 Naceanceno KS, House SL, Asaro PV. Shared-task worklists improve clinical trial recruitment workflow in an academic emergency department. Appl Clin Inform 2021;12(02):293–300
4 Sen A, Ryan P, Goldstein A, et al. Assessing eligibility criteria generalizability and their correlations with adverse events using big data for EHRS and clinical trials. In Proceedings of the Data Science Learning and Applications to Biomedical and Health Sciences Conference (Big Data Workshop organized by New York Academy of Sciences; 74–79
5 Thadani SR, Weng C, Bigger JT, et al. Electronic screening improves efficiency in clinical trial recruitment. J Am Med Inform Assoc 2009;16(6):869–873
6 Weng C. Optimizing clinical research participant selection with informatics. Trends in pharmacological sciences 2015;36 (11):706–709
7 Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA 2007;297 (11):1233–1240
8 Janson M, Edlund G, Kressner U, et al. Analysis of patient selection and external validity in the Swedish contribution to the COLOR trial. Surg Endosc 2009;23(08):1764–1769
9 van der Aalst CM, van Iersel CA, van Klaveren RJ, et al. Generalisability of the results of the Dutch-Belgian randomised controlled lung cancer CT screening trial (NELSON): does self-selection play a role? Lung Cancer 2012;77(01):51–57
10 Bress AP, Tanner RM, Hess R, Colantonio LD, Shimbo D, Muntner P. Generalizability of SPRINT Results to the U.S. Adult Population. J Am Coll Cardiol 2016;67(05):463–472
11 Weng C, Li Y, Ryan P, et al. A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. Appl Clin Inform 2014;5(02):463–479
12 Sen A, Ryan P, Goldstein A, et al. Correlating eligibility criteria generalizability and adverse events using Big Data for patients and clinical trials. Annals of the New York Academy of Sciences 2017;1387 (01):34–43
13 Sen A, Chakrabarti S, Goldstein A, Wang S, Ryan PB, Weng C. GIST 2.0: a scalable multi-trait metric for quantifying population representativeness of individual clinical studies. J Biomed Inform 2016;63:325–336
14 Cahan A, Cahan S, Cimino JJ. Computer-aided assessment of the generalizability of clinical trial results. Int J Med Inform 2017; 99:60–66
15 Reich C, Ryan PB, Belenkaya R, et al. OHDSI Common Data Model v6.0 Specifications. Accessed 2019 at: https://github.com/OHDSI/CommonDataModel/wiki
16 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(02):239–250
17 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(04):294–305
18 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(05):507–513
19 Aronson AR. 2001Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In Proceedings of the AMIA Symposium (p. 17). American Medical Informatics Association. Accessed 2021 at: https://pubmed.ncbi.nlm.nih.gov/11825149/
20 Kury F, Butler A, Yuan C, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. Sci Data 2020;7(01):281
21 Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. Stud Health Technol Inform 2015; 216:574–578
22 Chang AX, Manning CD. 2012, May. Sutime: a library for recognizing and normalizing time expressions. In: LREC. European Language Resources Association (ELRA); vol. 2012;3735–3740. Accessed 2021 at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/284_Paper.pdf
23 Laffin LJ, Besser SA, Alenghat FJ. A data-zone scoring system to assess the generalizability of clinical trial results to individual patients. Eur J Prev Cardiol 2019;26(06):569–575
24 Chatterjee P, Cymberknop LJ, Armentano RL. Nonlinear systems in healthcare towards intelligent disease prediction. In: Nonlinear Systems-Theoretical Aspects and Recent Applications. IntechOpen; 2019
25 Awad M, Khanna R. Support vector regression. In: Efficient Learning Machines. Apress, BerkeleyCA67–80
26 Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One 2018;13(08):e0202344
27 Sun Y, Butler A, Lin F, et al. The COVID-19 trial finder. J Am Med Inform Assoc 2021;28(03):616–621
28 Kim JH, Ta CN, Liu C, et al. Towards clinical data-driven eligibility criteria optimization for interventional COVID-19 clinical trials. J Am Med Inform Assoc 2021;28(01):14–22
29 Al-Lawati JA. Diabetes mellitus: a local and global public health emergency!. Oman medical journal 2017;32(03):177–179
30 Sun Y, Butler A, Stewart LA, et al. Building an OMOP common data model-compliant annotated corpus for COVID-19 clinical trials. J Biomed Inform 2021;118:103790
31 Sen A, Goldstein A, Chakrabarti S, et al. The representativeness of eligible patients in type 2 diabetes trials: a case study using GIST 2.0. J Am Med Inform Assoc 2018;25(03):239–247
32 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, pp. 954–955). IEEE
33 Li X, Liu H, Kury F, et al. A Comparison between Human and NLP-based Annotation of Clinical Trial Eligibility Criteria Text Using The OMOP Common Data Model. In AMIA 2021 Virtual Informatics Summit; 394-403