Journal of the American Medical Informatics Association, 25(3), 2018, 239–247
doi: 10.1093/jamia/ocx091
Advance Access Publication Date: 13 September 2017
Research and Applications

Research and Applications

The representativeness of eligible patients in type 2 diabetes trials: a case study using GIST 2.0

Anando Sen,[1] Andrew Goldstein,[1] Shreya Chakrabarti,[1] Ning Shang,[1] Tian Kang,[1] Anil Yaman,[2] Patrick B Ryan,[1,3] and Chunhua Weng[1]

[1]Department of Biomedical Informatics, Columbia University, New York, NY, USA, [2]Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands, and [3]Janssen Research and Development, Titusville, NJ, USA

Corresponding Author: Chunhua Weng, 622 W 168th Street, PH-20-407, New York, NY 10032, USA. Email: cw2384@cumc.columbia.edu

Received 9 May 2017; Revised 23 June 2017; Accepted 8 August 2017

ABSTRACT

Objective: The population representativeness of a clinical study is influenced by how real-world patients qualify for the study. We analyze the representativeness of eligible patients for multiple type 2 diabetes trials and the relationship between representativeness and other trial characteristics.

Methods: Sixty-nine study traits available in the electronic health record data for 2034 patients with type 2 diabetes were used to profile the target patients for type 2 diabetes trials. A set of 1691 type 2 diabetes trials was identified from ClinicalTrials.gov, and their population representativeness was calculated using the published Generalizability Index of Study Traits 2.0 metric. The relationships between population representativeness and number of traits and between trial duration and trial metadata were statistically analyzed. A focused analysis with only phase 2 and 3 interventional trials was also conducted.

Results: A total of 869 of 1691 trials (51.4%) and 412 of 776 phase 2 and 3 interventional trials (53.1%) had a population representativeness of <5%. The overall representativeness was significantly correlated with the representativeness of the Hba1c criterion. The greater the number of criteria or the shorter the trial, the less the representativeness. Among the trial metadata, phase, recruitment status, and start year were found to have a statistically significant effect on population representativeness. For phase 2 and 3 interventional trials, only start year was significantly associated with representativeness.

Conclusions: Our study quantified the representativeness of multiple type 2 diabetes trials. The common low representativeness of type 2 diabetes trials could be attributed to specific study design requirements of trials or safety concerns. Rather than criticizing the low representativeness, we contribute a method for increasing the transparency of the representativeness of clinical trials.

Keywords: population representativeness, clinical trials, eligibility criteria, metadata analysis

INTRODUCTION

Randomized controlled trials are the gold standard for generating medical evidence. Each trial has a set of eligibility criteria that defines a study population. Patients who satisfy these criteria are deemed eligible for the study. The representativeness of the eligible patients (referred to collectively as the study population)[1–3] influences the applicability and generalizability of the study results. Underrepresented population subgroups may suffer from unexpected postmarketing adverse events when the study interventions are applied to them.[4,5]

One approach to estimate the population representativeness of a study is to measure what percentage of real-world patients are eligible for the study.[6] Eligibility criteria of clinical studies have been criticized for being restrictive and complex and lacking clarity.[7,8] However, resources to assist clinical investigators with optimizing the representativeness of eligibility criteria design are very scarce. Reuse of eligibility criteria from previous trials is a common practice.[9] Alternatives include empirical knowledge, which can be gained through iterative trial and error and applied to new study designs or through frequent ad hoc protocol amendments to adjust recruitment needs.[10] Proactive data-driven decision aids for optimizing eligibility criteria design are unavailable but much needed.[11]

Previous studies have demonstrated[9,12] that the lack of population representativeness is not restricted to individual studies, but is also observed generally within the research community, eg, among multiple studies across various disease domains. Studies have performed collective assessments of clinical trial population representativeness using single (or a small set of) traits. For example, Somerson et al.[13] demonstrated a lack of racial diversity in a set of 158 orthopedic clinical trials. In particular, African American and Hispanic populations were shown to be underrepresented. Schoenmaker et al.[14] showed that dementia trial patients are older than real-world dementia patients. Hoertel et al.[15] showed that over half of the bipolar disorder patients from a nationally representative sample of 43 093 patients would fail at least one eligibility criterion in 87 bipolar depression and acute mania trials.

We previously designed the Generalizability Index of Study Traits (GIST) metric to measure the a priori representativeness of eligibility criteria of related trials,[6] and extended GIST to GIST 2.0.[16] The GIST 2.0 methodology differs from GIST and other representativeness metrics, such as propensity scores,[17,18] due to its explicit modeling of trait dependencies. We previously used GIST to compute the collective population representativeness of multiple related type 2 diabetes clinical studies using multiple study traits.[2] In this paper, we used GIST 2.0 to quantify the representativeness of the eligibility criteria of individual type 2 diabetes mellitus trials downloaded from ClinicalTrials.gov and studied the relationships between population representativeness and other trial characteristics, such as duration, phase, and so on.

Our study aims to address the following 5 research questions: (1) How representative are the study eligibility criteria of individual type 2 diabetes trials as measured by GIST 2.0? (2) How does the number of criteria included for GIST 2.0 calculation affect the study's population representativeness? (3) Which eligibility criterion has a relatively larger effect on a trial's population representativeness? (4) How does the duration of a trial relate to its population representativeness? (5) How is a clinical trial's population representativeness related to study metadata such as study phase, intervention type, etc.? The answers to these questions can potentially inform future methods for optimizing the design of eligibility criteria.[11]

METHODS

Glossary

The following terms are used frequently in this paper.

1. Trait: observable patient characteristics (eg, diagnoses, laboratory tests). In a clinical trial, each eligibility criterion is a rule on one or more traits (eg, glucose >126 mg/dL, no prior stroke). The corresponding traits are therefore referred to as eligibility traits.
2. Target population: the subset of all patients to whom the study results are applicable.
3. Study population: the subset of patients within the target population who satisfy all eligibility criteria of the trial.

Before calculating population representativeness, we performed 3 preprocessing steps: selection of trials, selection of traits, and definition of the target population, as shown in Figure 1.

Trial and trait selection

In this study, we considered the trials testing hypotheses on one condition, type 2 diabetes mellitus. The rationale was to compare the population representativeness across all the trials with the same target population, ie, patients with type 2 diabetes mellitus (the definition of which is described in the next section). Trials investigating multiple conditions (eg, type 2 diabetes with chronic kidney disease) were excluded, as they may have smaller target populations. At the point of the study (March 2016), there were 220 842 trials listed in ClinicalTrials.gov, 4576 of which had type 2 diabetes mellitus as a condition. After excluding 1736 trials with multiple conditions, 2840 trials with type 2 diabetes as the sole condition were retained. The eligibility criteria of these trials were parsed using a published parser, Eligibility Criteria Extraction and Representation (EliXR).[19] This software recognizes Unified Medical Language System (UMLS) concepts in the free text of eligibility criteria.

A frequency table was generated for the UMLS concept identifiers and further consolidated to account for synonyms (eg, "cancer" and "malignant tumors") and concepts with multiple identifiers (eg, 2 UMLS identifiers for "pregnancy" are C3484365 and C3539106). Only the concepts (hereafter referred to as traits) with >120 occurrences were included for representativeness analysis. Traits that were unavailable in electronic health records (EHRs), such as informed consent and participation in other clinical trials, were excluded from analysis.

Quantitative traits that were prevalent in <5% of patients in the EHR (eg, C-peptide) were also excluded. The remaining quantitative traits were parsed using Valx[20] to extract the upper and lower limits of each trait. We eventually selected 69 traits available in EHRs that are frequently used in type 2 diabetes trials, as listed in Table 1.

The 1763 trials that included criteria for at least 2 distinctive traits were included for further analysis. All numerical traits were manually reviewed after automated parsing to detect parsing errors. The 2 most common error types were missing values for both upper and lower limits, and lower limits higher than upper limits. A total of 72 trials had such errors and were excluded from our trial list. Subsequently, we included 1691 trials for analysis.

We also conducted a focused analysis that included only phase 2 and 3 interventional trials. Trials of phase 0 and 1 are primarily meant for dose-ranging and safety evaluations, often on healthy volunteers. This can result in their eligibility criteria being less restrictive. Phase 4 and observational trials are meant for long-term effects. Hence, phase 2 and 3 interventional trials are the ones that determine the effectiveness of an intervention on patients.

Target population definition

We first extracted a random sample of 30 000 patients from the 4.5 million patients in the Columbia University Medical Center Clinical Data Warehouse. The time frame of EHRs within the Clinical Data Warehouse ranged from the 1980s to the point of data extraction, August 2015. A total of 5273 type 2 diabetes mellitus patients were further identified from this sample as (1) having an International Classification of Diseases, Ninth Revision (ICD9) code for type 2

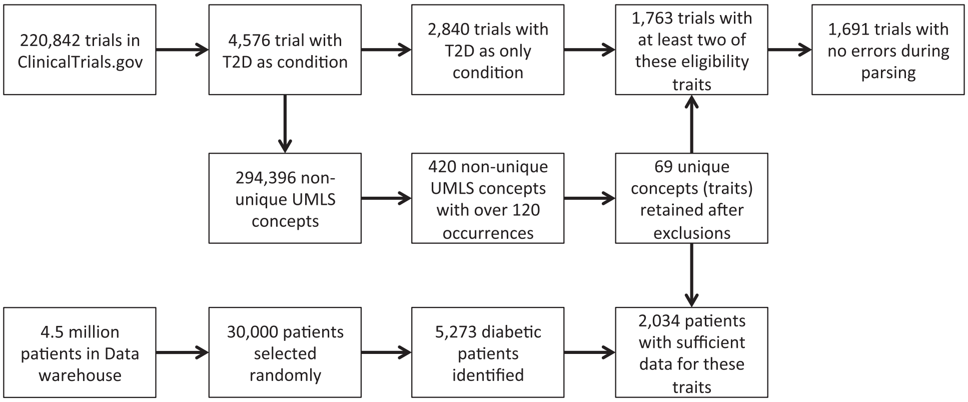
220,842 trials in ClinicalTrials.gov
4,576 trial with T2D as condition
2,840 trials with T2D as only condition
1,763 trials with at least two of these eligibility traits
1,691 trials with no errors during parsing
294,396 non-unique UMLS concepts
420 non-unique UMLS concepts with over 120 occurrences
69 unique concepts (traits) retained after exclusions
4.5 million patients in Data warehouse
30,000 patients selected randomly
5,273 diabetic patients identified
2,034 patients with sufficient data for these traits

Figure 1. Summary of the preprocessing steps: trial selection (top row), trait selection (second row), and definition of target population (third row).

Table 1. List of frequently used study traits in type 2 diabetes trials

Demographic	Age	Gender			
Medications	Metformin	Glucagon	Beta-blockers	Sulfonylurea	Thiazolidinediones
General Conditions	Drug abuse Tobacco abuse	Pregnant	Alcohol abuse	Breastfeeding	Substance abuse
Laboratory	Hba1c Bilirubin eGFR	AST ALT LDL	HDL Fasting glucose	Triglycerides Hemoglobin	Total cholesterol Creatinine
Procedures	Kidney transplant Dialysis	Surgery	Major surgery	Weight loss surgery	Coronary bypass surgery
Diagnoses	Pancreatitis Neuropathy Hyperglycemia Heart failure Cerebral stroke Cardiovascular disease Cerebrovascular disease Gastrointestinal disease	Anemia Angina Cancer HIV Hepatitis C Myocardial infarction Gestational diabetes Basal cell skin cancer	Type 1 diabetes Gastroparesis Renal disease Hepatitis B Thyroid cancer Diabetic ketoacidosis Hematological disorder Peripheral arterial disease	Pre-diabetes Hypertension Hypoglycemia Arrhythmia Retinopathy Proliferative retinopathy Inflammatory bowel disease	Liver cirrhosis Pulmonary disease Endocrine disease Chronic pancreatitis Liver disease Transient ischemic attack Coronary artery disease

AST: aspartate aminotransferase; ALT: alanine transaminase; HDL: high-density lipoprotein; LDL: low-density lipoprotein; eGFR: estimated glomerular filtration rate; HIV: human immunodeficiency virus.

diabetes or (2) satisfying the World Health Organization (WHO) criteria for diabetes, ie, fasting glucose >126 mg/dL or Hba1c >6.5%.[21] When phenotyping with the WHO criteria, patients with type 1 diabetes ICD9 codes were excluded. This phenotyping method results in both incident and prevalent cases of type 2 diabetes being selected. Several previous studies called for the use of phenotyping algorithms to identify diabetic patients from EHRs.[22–24] In particular, Spratt et al. compared various phenotyping algorithms for type 2 diabetes. For the ICD9 condition, the sensitivity and specificity were 0.91 and 0.97, respectively. For the Hba1c condition (without the use of fasting glucose), the corresponding values were 0.69 and 0.99.

For all categorical traits except gender (ie, medications, procedure, general conditions, and diagnoses), we assigned a positive truth value to a patient if and only if we found definite evidence of the trait (eg, ICD9 codes, medication orders) in the patient's structured EHR data. Otherwise, we marked negative. For quantitative traits, we calculated the median of all readings. Predictive mean-matching multiple imputation[25] was used to fill in missing quantitative values by using the method described by Rubin,[26] while ensuring that that no trait had >5% missing values. The target population (with imputed values) consisted of 2034 diabetes patients.

The GIST 2.0 metric

The previously published GIST 2.0 methodology[16] computes a multiple-trait GIST (mGIST) score for the entire study and one single-trait GIST (sGIST) score for each eligibility trait. The mGIST score (when computed using all traits) approximates the fraction of the target population that would be eligible for the study. When only a subset of the eligibility traits is used for the mGIST calculation, the calculated score is an approximation to the true mGIST score. Similarly, the sGIST score of a particular eligibility trait approximates the fraction of the target population that would satisfy the eligibility criterion for that trait. The GIST 2.0 algorithm inputs the EHR data of a target population and the eligibility criteria, which are used to profile the study population, and outputs the degree of overlap between the target population and the study population.

GIST 2.0 accounts for the interdependence between the various traits with a Gaussian kernel-based regression hypersurface. The fitting of the hypersurface accommodates both quantitative and categorical variables. The significance of each trait is modeled as an inverse relation to the stringency of its eligibility criterion. Moreover, GIST 2.0 uses a weighting scheme that minimizes the effects of outliers. Rigorous mathematical details on GIST 2.0 (with proofs) can be found in Sen et al.,[16] and a less technical schematic version is described in Sen et al.[5] A brief summary of the methodology is provided in the Supplementary Appendix.

Statistical analyses

From the trial description in ClinicalTrials.gov, we extracted all available attributes of the trial. The trial duration was calculated as the difference between the start date and end date (wherever both of these were available) and further categorized into 1-year intervals. Only the trials with end dates before the point of study, March 2016 (ie, confirmed end dates as opposed to prospective end dates), were considered for the duration analysis. Trials longer than 5 years were binned into the category "5 or higher." Statistical differences between these categories were analyzed with a one-way analysis of variance (ANOVA).

Among all the study metadata, we analyzed the effects of study phase, study type, recruitment status, start year, and funding source on the mGIST scores. Study phases included phases 0 through 4. Funding agencies included industry, US federal agencies, and others. Study type was either observational or interventional. Recruitment status was one of the following: recruiting, active but not recruiting, completed, enrolling by invitation, not yet recruiting, suspended, terminated, and withdrawn. Since several of these groups had very few trials, they were aggregated into 2 categories: No further recruitment – completed, terminated, active but not recruiting, and withdrawn; and further recruitment possible – recruiting, enrolling by invitation, not yet enrolling, and suspended. Start years were from 2001 through 2016. Since only 11 trials in 2016 were included at the point of the study, they were combined with the 2015 trials within the category "2015 or later." Due to the smaller number of phase 2 and 3 interventional trials, the start date effect was categorized in 2-year intervals (ie, 2001–2002, 2003–2004, etc.), and the last 2 categories were combined to form the single category "2013 or later."

These effects were statistically analyzed by a 5-way ANOVA studying all the effects together. Only those trials that had information for all 5 effects were included. For recruitment status, an initial one-way ANOVA studied the effects of subcategories within each category on population representativeness to ensure that there were no discrepancies within the categories. To study 2-way interaction effects, the 5-way ANOVA was repeated after excluding phase 0 trials (due to their low counts). Statistical significance for all tests was calculated at the 0.05 level.

RESULTS

Although we cannot share proprietary EHR data, demonstrations of our codes for computing GIST 2.0 scores using synthetic data are available at Github: https://github.com/anandosen/pop_rep. We organize our results as answers to the research questions posed above.

How representative are the study populations of diabetes trials as measured by GIST 2.0?

Figure 2 (A) shows the histogram of the mGIST scores for all 1691 trials and for 776 phase 2 and 3 interventional trials, respectively. An exponential decay of the scores is visible for both groups. About a quarter of the trials have mGIST scores <0.01 (414 of 1691 in the general case and 202 of 776 for phase 2 or 3 interventional trials). The median scores for the 2 groups were 0.048 and 0.045, respectively, as marked in Figure 2 (B), where the corresponding cumulative distribution functions are shown. The cumulative distribution function for an mGIST score (on the horizontal axis) is the fraction of trials that have a lower or equal mGIST score. The median points lie between 0.04 and 0.05 on the horizontal axis in both cases. In other words, >50% of the trials (869 of 1691 and 412 of 776, respectively) have representativeness scores <0.05.

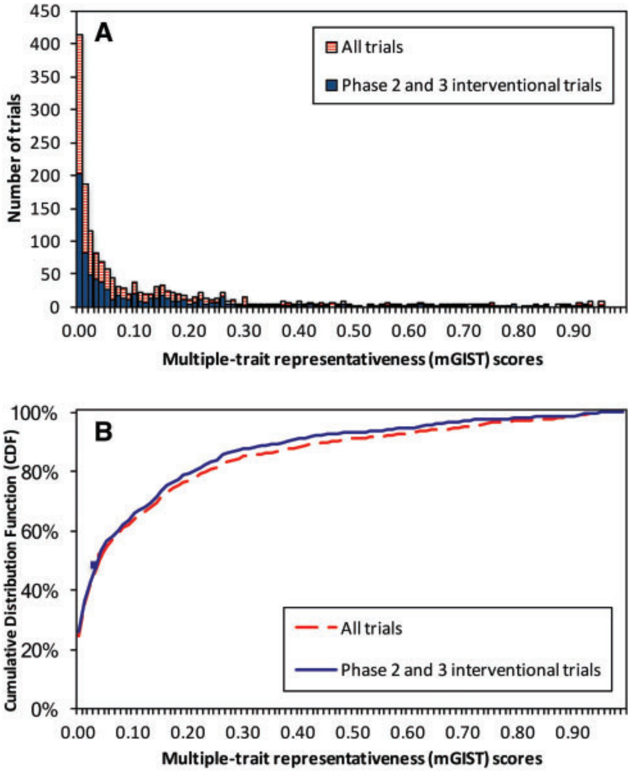

Figure 2. (**A**) Histogram of the multitrait representativeness scores (mGIST) for type 2 diabetes trials. (**B**) The corresponding cumulative distribution function, with the position of the median marked.

We used sGIST scores to identify the traits that excluded the most patients. We present only the case of all trials, as results were similar for phase 2 and 3 interventional trials. For individual traits, median sGIST calculations included only those trials where the trait was a part of the eligibility criteria. Cardiovascular diseases (the entire spectrum as opposed to a particular disease) had the lowest median sGIST at 0.08. This was followed by beta-blockers and thiazolidinediones (both median sGISTs ~0.16), hypertension (0.20), and gastrointestinal diseases (0.27). All of these are common medications or diseases.

In 340 of the 1691 trials, the mGIST score was computed using all of their study traits. Hence, for these trials, the calculated mGIST score was equal to the true mGIST score. In the other cases, mGIST was calculated using a subset of the eligibility traits available in EHRs and hence was an upper bound[16] to the true mGIST score. To evaluate how the fraction of traits used for mGIST calculations affected the distribution, we regenerated Figure 2 (A) for the cases: (1) trials where >25% of the traits were used for mGIST calculations; (2) trials where >50% of the traits were used for mGIST calculations; (3) trials where >75% of the traits were used for mGIST calculations; and (4) trials where 100% of the traits were used for mGIST calculations. This analysis is again presented only for the entire set of 1691 trials, as results were similar for phase 2 and 3 interventional trials.

To account for the different number of trials in each of these cases (1352, 905, 516, and 340, respectively), we present the relative

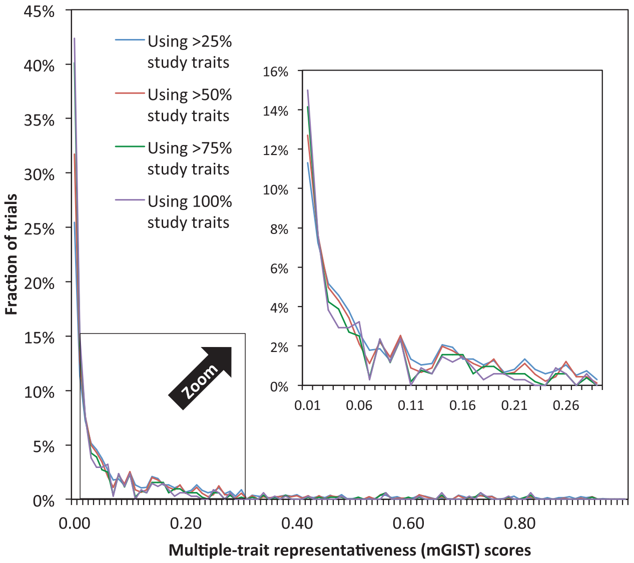

Figure 3. Relative histogram for multitrait representativeness scores (mGIST) of type 2 diabetes trials for different fractions of traits used in the representativeness calculations.

histogram (or the sample probability density function). In a relative histogram, the frequency of a particular bin is normalized between 0 and 1 by dividing the total number of trials for that case. Figure 3 presents these relative histograms. As can be seen, the distributions are very similar. A one-way ANOVA testing the effect of fraction of traits used for mGIST calculations found no significant differences between these 4 cases ($P \sim 1.0$). The medians for these categories were 0.041, 0.027, 0.014, and 0.013, respectively.

How does the number of traits selected for GIST calculation affect the study's population representativeness?

In the formulation of GIST 2.0, we proved that for a trial, the addition of an eligibility criterion would lower its mGIST score.[16] Hence we hypothesize that for a collection of trials, the mGIST score would decay with the number of eligibility traits used for the mGIST calculation. For the both the 1691 trials and the phase 2 and 3 interventional trials, the trials were grouped by number of eligibility criteria, and mean mGIST scores (for each of the categories) were calculated. This is shown in Figure 4. As expected, the mGIST scores decrease as the number of eligibility criteria increases. The decays are exponential and are confirmed by statistically significant Spearman's correlation coefficients of −0.98 and −0.96, respectively. In fact, when trials have >20 eligibility criteria, the mGIST score is close to zero.

Which criterion has a relatively larger effect on a study's population representativeness?

We computed the sGIST scores for each trait and identified Hba1c to have the highest correlation with mGIST. In the general case, among the 1691 trials, 1324 had eligibility criteria for Hba1c. For these trials, the Pearson's correlation coefficient between the Hba1c sGIST score and the mGIST score was 0.44, with $P < .01$. This implies a moderate but statistically significant correlation. Figure 5 shows the relative histograms for the mGIST and the Hba1c sGIST. While the mGIST has a one-sided peak near zero, the peak for Hba1c is to its right at 0.3. This is in agreement with the property of GIST scores, which states that sGIST scores are always higher than mGIST scores.[16] A difference between the curves is that about 15% of the trials have very high sGISTs (>0.9). These are trials where Hba1c is used as an exclusion criterion (eg, Hba1c >14%). The mGIST is rarely this high. For phase 2 and 3 interventional trials (634 in total), the results were almost identical. The Pearson's correlation coefficient was again 0.44.

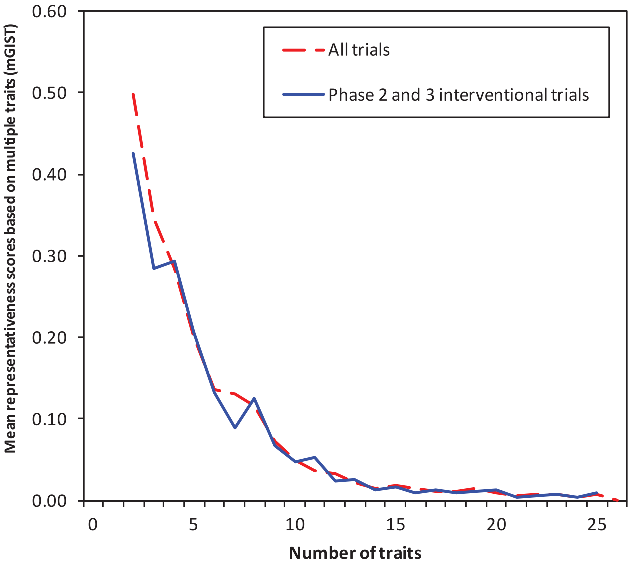

Figure 4. Relationship between number of eligibility traits and mean representativeness score based on multiple traits (mGIST).

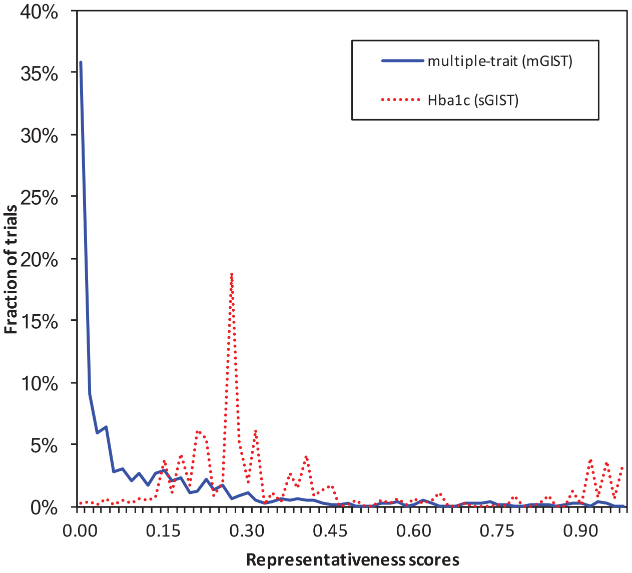

Figure 5. Relative histogram for multitrait representativeness scores (mGIST) and single-trait representativeness scores for Hba1c (sGIST) for the 1324 trials where Hba1c was an eligibility trait.

How does the duration of a trial correlate with its population representativeness?

The effect of duration on population representativeness is shown in Table 2. A total of 1424 trials had start and end date information, with the end date before March 2016. The one-way ANOVA with

Table 2. Summary of duration effect on mGIST score

Trial Duration (years)	Number of Trials (all)	Mean mGIST	Number of Phase 2 and 3 Interventional Trials	Mean mGIST
0–1	433	0.14	176	0.11
1–2	553	0.13	312	0.11
2–3	231	0.16	103	0.15
3–4	97	0.15	30	0.14
4–5	60	0.16	19	0.21
≥5	50	0.21	19	0.19

duration as the effect resulted in a statistically significant *P*-value of .0392. The population representativeness was positively correlated with the trial duration, as demonstrated by a high Spearman's correlation coefficient of 0.88. Of the above trials, 659 were phase 2 and 3 interventional trials. For these, the *P*-value from the one-way ANOVA was significant at .0481, and the Spearman's correlation coefficient was 0.82.

How is a clinical trial's population representativeness related to study metadata?

Recall that the 5 effects we studied were study phase, funding agency, study type, recruitment status, and start year. When all trials were considered, each phase had a mean mGIST between 0.13 and 0.21. Of the 4 main phases, the mean mGIST decreased from 0.20 to 0.13 from phases 1 to 3 and then increased slightly to 0.15 for phase 4 (Table 3). The ANOVA showed phase as a statistically significant factor for mGIST, with a *P*-value of .0017. Study type can be either interventional or observational. Nearly 95% of the studies were interventional (Table 3). Study type was found not to have a significant effect on mGIST, as both study types had similar mean mGISTs. A detailed description of the studies by recruitment status is also shown in Table 3. The initial one-way ANOVA for subcategories within the 2 categories found no significant intra-category differences. The corresponding *P*-values for the "further recruitment possible" and "no further recruitment" categories were 0.4964 and 0.6372, respectively. Trials that had further recruitment possible had a higher mean mGIST. This difference was statistically significant, with a *P*-value of 0.0198. When grouped by funding agency, industrial funding had the maximum number of trials but also a lower mGIST score than the other groups (Table 3). However, the difference in the mGIST scores between different metadata was not statistically significant. Finally, we tested the effect of start year of a trial on mGIST. Table 3 also shows the mGIST for every year from 2001 through 2015. The number of trials in each year changed from 21 in 2001 to 179 in 2009. The mGISTs show a somewhat decreasing pattern, as confirmed by a moderate but significant Spearman's correlation coefficient of −0.48. Start year turned out to be a significant effect on mGIST, as the ANOVA yielded a *P*-value of 0.0162. Among the interaction effects, the phase–funding agency and the recruitment status–funding agency pairs were found to be statistically significant, with *P*-values of 0.0455 and 0.0366, respectively.

The corresponding summary for the phase 2 and 3 interventional trials is shown in Table 4. A total of 758 trials having all metadata information were included in this analysis. As is evident from Table 3, phases 2 and 3 have very similar mean mGISTs, and hence their representativeness was not statistically different. When grouped by recruitment status, the "further recruitment possible" category was reduced to just 87 trials, which led to the loss of statistical significance (*P*-value = .20). The start date effect remained statistically significant (*P*-value = .0020). The decreasing trend was still observed, except for the last category. This category contained only 32 trials due to the exclusion of a large number of phase 4 trials. The effect of funding agency continued to remain statistically insignificant.

Table 3. Summary of metadata effects. Each section corresponds to an effect and is stratified by that effect.

Effect	Number of Trials	Mean mGIST	*P*-value
Overall	1691	0.149	
Phase			**.0017**
Phase 0	11	0.21	
Phase 1	160	0.20	
Phase 2	253	0.13	
Phase 3	525	0.13	
Phase 4	373	0.15	
Type			.1558
Interventional	1607	0.14	
Observational	82	0.15	
Recruitment Status			**.0198**
Further recruitment possible	335	0.16	
No further recruitment	1355	0.14	
Funding Agency			.8965
Industry	1175	0.14	
Federal agency	59	0.15	
Other	456	0.16	
Start Year			**.0162**
2001	21	0.21	
2002	40	0.17	
2003	68	0.14	
2004	56	0.13	
2005	69	0.20	
2006	109	0.15	
2007	110	0.13	
2008	147	0.16	
2009	179	0.14	
2010	172	0.11	
2011	156	0.18	
2012	131	0.12	
2013	126	0.14	
2014	120	0.13	
2015 or later	151	0.14	

Significant *P*-values are in bold.

DISCUSSION

Implications of the metadata findings

Our study is the first of its kind to replicate findings from the literature about the lack of representativeness in clinical trial eligibility criteria using EHR data. Our finding also helps researchers and the public understand why clinical trial recruitment is hard, because only a small portion of real-world patients are eligible for these trials. The decrease in population representativeness in recent trials can be attributed to a larger number of exclusion criteria in trials, possibly due to stricter safety regulations in recent years. Yao et al.[27] describe the trends in clinical trial patient safety as "having evolved with increased requirements for risk management plans, risk evaluation and minimization strategies." The International Conference on Harmonization maintains detailed safety guidelines for clinical studies.[28] These guidelines have been updated regularly since 1995 and have been enforced as laws in several countries. The decrease in population representativeness from phase 1 through phase 3 trials can initially seem

Table 4. Summary of metadata effects for phase 2 and 3 interventional trials. Each section corresponds to an effect and is stratified by that effect.

Effect	Number of Trials	Mean mGIST	*P*-value
Overall	758	0.129	
Phase			.8034
Phase 2	249	0.13	
Phase 3	509	0.13	
Recruitment Status			.7387
Further recruitment possible	87	0.14	
No further recruitment	671	0.13	
Funding Agency			.1292
Industry	665	0.13	
Federal agency	10	0.07	
Other	83	0.16	
Start Year			**.0020**
2001–2002	42	0.20	
2003–2004	88	0.15	
2005–2006	116	0.14	
2007–2008	138	0.12	
2009–2010	213	0.11	
2011–2012	129	0.09	
2013 or later	32	0.20	

Significant *P*-values are in bold.

counterintuitive, as study populations enlarge from phase 1 to phase 3, ie, phase 1 trials typically have <30 patients, while phase 3 trials can have several thousands.[29] However, phase 1 trials are generally meant for safety evaluations[29] and phase 3 trials are meant to have more stringent eligibility criteria, as all subpopulations deemed "unsafe" are excluded. Phase 4 trials, which are meant for the wider community, tend to have less restrictive criteria than phase 3. Statistical significance was lost in the case of phase 2 and 3 interventional trials. This further confirms that the significant difference due to phase was coming from phase 1 to 4 trials, as their objectives are different. This meta-analysis can be an important guide during trial design. By computing the a priori population representativeness during study design, designers can make revisions to the criteria or present justifications for the lack of population representativeness.

Broader implications

The most striking result in our analysis is that over half of the diabetes trials exclude >95% of the target population by design. George[30] discussed the general reasons for restrictive eligibility criteria. Though presented for cancer trials, the reasons apply to multiple medical conditions generally. The reasons can be scientific or due to safety concerns. Scientific reasons include restricting the study participants to a subset of the target population. For example, the SPRINT[31] study only defines the target population as high-risk hypertension patients (as opposed to all hypertension patients). This leads to complex eligibility criteria, which are often blamed for low representativeness.[32,33] The risk of adverse events is the primary safety consideration. Statler et al.[34] showed that this consideration often leads to overly restrictive eligibility criteria, though a direct association could not be established. Among other reasons, geographical constraints (eg, ease of access) can lower representativeness.[35] Though we did not include geographic characteristics in this study, it is well known that study traits are correlated with geographic location.[36]

In addition, we found that trials with a greater number of eligibility criteria lead to lower population representativeness. An implication of this finding is that clinical study recruitment is hard by design because of the large number of eligibility criteria,[33,37] which automatically reduces the representativeness of the study population. The mentioned article[37] also shows that the number of eligibility criteria rose between 2000 and 2012. This offers further evidence for the temporal decline of representativeness among trials conducted over time.

We have shown that long-duration trials have higher representativeness. Long-duration trials typically include a follow-up period when patients are monitored over several years. There may be significant loss of contact with patients during this follow-up period. Bianchi[38] suggested adequate and proper enrollment during the recruitment process as one way of overcoming loss of patients during the follow-up period. Having relatively less-restrictive eligibility (hence higher representativeness) criteria within safety constraints can potentially aid in securing adequate enrollment. Moreover, several long-duration trials are pragmatic trials whose recruitment goal is to have participants similar to the patients who would receive the intervention as part of usual care[39] (unlike regular trials, which provide optimal care). Hence, eligibility criteria are kept to a minimum[40] and population representativeness is higher.

LIMITATIONS

Our study has certain limitations. Though we identified trials studying only type 2 diabetes, certain problems remain in isolating these trials. Some trials state only "type 2 diabetes" in their conditions field, but they may be studying additional conditions as well. For example, the NCT01043029 trial studies kidney disease in addition to diabetes and requires moderately impaired renal function (eGFR 30–59 ml/min/1.73 m^2).[41] This trial's representativeness should ideally be computed with a smaller target population. Similarly, some trials (eg, NCT02188186) might be studying only severe cases of type 2 diabetes and should not enroll all diabetes patients.

The use of EHRs to define target populations is a common practice but can introduce certain biases and inaccuracies. Several studies have shown significant differences between certain traits[9,42] of real-world patients and patients receiving hospital care reflected in EHRs. ICD9 codes for within EHRs are primarily meant for billing purposes, and the process of assigning ICD9 codes can have several sources of error.[43] Further, EHRs often do not record traits important for clinical trials. For example, even if not mentioned in the eligibility criteria, informed consent is always required for a clinical trial. Hence, it is virtually impossible to consider a complete set of traits in the computation of mGIST. As mentioned above, this implies that the computed score is actually an upper bound of the true mGIST score. However, with a median of 0.048 (lowered to 0.013 when all of the traits were considered), this score can still provide valuable information. An example is a correlation between the distribution of mGIST and sGIST for Hba1c, which shows that Hba1c is one of the most important traits in determining the population representativeness of diabetes trials. This is in fact true, as Hba1c is one of the defining traits for type 2 diabetes.[21]

CONCLUSIONS

More than 50% of type 2 diabetes trials' eligibility criteria exclude >95% of type 2 diabetes patients. Study phase, recruitment status, and trial start year significantly affect a trial's population representativeness. Further research on clinical trial population representativeness and the optimal balance between intervention outreach and patient safety is warranted, which is a complex problem that needs joint efforts from clinical research designers and practitioners.

FUNDING

This project was supported by National Library of Medicine grant R01LM009886 (PI: Weng).

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

AS performed the majority of the data extraction and data analysis, designed the methodology, and wrote the manuscript. AG was the medical expert and supervised the trial and trait selection process. SC, NS, and PR provided input for the methodology and reviewed the manuscript. TK and AY performed part of the data extraction that involved EliXR and Valx. CW designed and supervised the research and edited the manuscript substantially.

SUPPLEMENTARY MATERIAL

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

ACKNOWLEDGMENTS

The authors would like to thank Prof Bruce Levin for reviewing the paper and providing important feedback on the statistical analysis and impact evaluation. Dr Maurine Tong and Dr Jung Hoon Son provided useful input during the course of this project.

REFERENCES

1. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials*. 2015;16:1–14.
2. He Z, Ryan P, Hoxha J, Wang S, Carini S, Sim I, Weng C. Multivariate analysis of the population representativeness of related clinical studies. *J Biomed Inform*. 2016;60:66–76.
3. Wang TS, Hellkamp AS, Patel CB, Ezekowitz JA, Fonarow GC, Hernandez AF. Representativeness of RELAX-AHF clinical trial population in acute heart failure. *Circ Cardiovasc Qual Outcomes*. 2014;7:259–68.
4. Masoudi FA, Havranek EP, Wolfe P, *et al*. Most hospitalized older persons do not meet the enrollment criteria for clinical trials in heart failure. *Am Heart J*. 2003;146:250–57.
5. Sen A, Ryan PB, Goldstein A, Chakrabarti S, Wang S, Koski E, Weng C. Correlating eligibility criteria generalizability and adverse events using Big Data for patients and clinical trials. *Ann NY Acad Sci*. 2017;1387:34–43.
6. 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;5:463–79.
7. Musen MA, Rohn JA, Fagan LM, Shortliffe EH. Knowledge engineering for a clinical trial advice system: Uncovering errors in protocol specification. *Bull Cancer*. 1987;74:291–96.
8. Ross J, Tu S, Carini S, Sim I. Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits Transl Sci Proc* 2010;46–50.
9. Hao T, Rusanov A, Boland MR, Weng C. Clustering clinical trials with similar eligibility criteria features. *J Biomed Inform*. 2014;52:112–20.
10. Rubin DL, Gennari J, Musen MA. Knowledge representation and tool support for critiquing clinical trial protocols. *Proc AMIA Annu Symp*. 2000;724–28.
11. Weng, C. Optimizing clinical research participant selection with informatics. *Trends Pharmacol Sci*. 2015;36:706–09.
12. He Z, Ryan P, Hoxha J, Wang S, Carini S, Sim I, Weng C. Multivariate analysis of the population representativeness of related clinical studies. 2016;60:67–76.
13. Somerson JS, Bhandari M, Vaughan CT, Smith CS, Zelle BA. Lack of diversity in orthopaedic trials conducted in the United States. *J Bone Joint Surg Am*. 2014;96:e56.
14. Schoenmaker N, Van Gool WA. The age gap between patients in clinical studies and in the general population: a pitfall for dementia research. *Lancet Neurol*. 2004;3:627–30.
15. Hoertel N, Strat Y Le, Lavaud P, Dubertret C, Limosin F. Generalizability of clinical trial results for bipolar disorder to community samples. *J Clin Psychiatry*. 2013;74:265–70.
16. 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;63:325–36.
17. Pressler TR, Kaizar EE. The use of propensity scores and observational data to estimate randomized controlled trial generalizability bias. *Stat Med*. 2013;32:3552–68.
18. Greenhouse JB, Kaizar EE, Kelleher K, Seltman H, Gardner W. Generalizing from clinical trial data: A case study. The risk of suicidality among pediatric antidepressant users. *Stat Med*. 2008;27:1801–13.
19. Weng C, Wu X, Luo Z, Boland MR, Theodoratos D, Johnson SB. EliXR: an approach to eligibility criteria extraction and representation. *J Am Med Inform Assoc*. 2011;18 (Suppl 1):i116–24.
20. Hao T, Liu H, Weng C. Valx: a system for extracting and structuring numeric lab test comparison statements from text. *Methods Inf Med*. 2016;55:266–75.
21. Vijan S. Type 2 diabetes. *Ann Intern Med*. 2010;152:ITC315.
22. Richesson RL, Hammond WE, Nahm M, *et al*. Electronic health records based phenotyping in next-generation clinical trials: a perspective from the NIH Health Care Systems Collaboratory. *J Am Med Inform Assoc*. 2013;20:e226–31.
23. Spratt SE, Pereira K, Granger BB, *et al*. Assessing electronic health record phenotypes against gold-standard diagnostic criteria for diabetes mellitus. *J Am Med Inform Assoc*. 2016;42:e121–28.
24. Anderson AE, Kerr WT, Thames A, Li T, Xiao J, Cohen MS. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study. *J Biomed Inform*. 2016;60:162–68.
25. Morris TP, White IR, Royston P. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Med Res Methodol*. 2014;14:75.
26. Rubin DB. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons; 1987.
27. Yao B, Zhu L, Jiang Q, Xia HA. Safety monitoring in clinical trials. *Pharmaceutics*. 2013;5:94–106.
28. International Council for Harmonisation. *Safety Guidlines*. 2016. http://www.ich.org/products/guidelines/safety/article/safety-guidelines.html.
29. Cancer.Net Editorial Board. *Phases of Clinical Trials*. 2015.
30. George SL. Reducing patient eligibility criteria in cancer clinical trials. *J Clin Oncol*. 1996;14:1364–70.
31. Bress AP, Tanner RM, Hess R, Colantonio LD, Shimbo D, Muntner P. Generalizability of results from the Systolic Blood Pressure Intervention Trial (SPRINT) to the US adult population. *J Am Coll Cardiol*. 2016;67:463–72.
32. Kadam RA, Borde SU, Madas SA, Salvi SS, Limaye SS. Challenges in recruitment and retention of clinical trial subjects. *Perspect Clin Res*. 2016;7:137–43.
33. Gelling L. *Facing the Challenges of Recruitment to Clinical Trials – Clinfield*. 2016. http://clinfield.com/2016/06/recruitment/.

34. Statler A, Radivoyevitch T, Siebenaller C, *et al*. The relationship between eligibility criteria and adverse events in randomized controlled trials of hematologic malignancies. *Leukemia*. 2016;31(8):1808–15.
35. Galsky MD, Stensland KD, McBride RB, *et al*. Geographic accessibility to clinical trials for advanced cancer in the United States. *JAMA Intern Med*. 2015;175:293.
36. Zaman MJ, Patel A, Chalmers J, *et al*. The effects of patient characteristics and geographical region on hospitalization in patients with Type 2 diabetes. *Diabet Med*. 2013;30:918–25.
37. Lopienski, K. Why do recruitment efforts fail to enroll enough patients? Nimblify, Inc. 2014. https://forteresearch.com/news/recruitment-efforts-fail-enroll-enough-patients/. Accessed June 20, 2017.
38. Bianchi A. Patient recruitment driving length and cost of oncology clinical trials. *Int Pharm Ind*. 2013;5:58–61.
39. Ford I, Norrie J. Pragmatic trials. *N Engl J Med*. 2016;375:454–63.
40. Patsopoulos NA. A pragmatic view on pragmatic trials. *Dialogues Clin Neurosci*. 2011;13:217.
41. Ruilope L, Hanefeld M, Lincoff AM, *et al*. Effects of the dual peroxisome proliferator-activated receptor-α/γ agonist aleglitazar on renal function in patients with stage 3 chronic kidney disease and type 2 diabetes: a Phase IIb, randomized study. *BMC Nephrol*. 2014;15:180.
42. Weiskopf NG, Rusanov A, Weng C. Sick patients have more data: the non-random completeness of electronic health records. *AMIA Annu Symp Proc*. 2013;1472–77.
43. O'Malley KJ, Cook KF, Price MD, Wildes KR, Hurdle JF, Ashton CM. Measuring diagnoses: ICD code accuracy. *Health Serv Res*. 2005;40:1620–39.