Simulation-based Evaluation of the Generalizability Index for Study Traits

Zhe He, PhD[1], Praveen Chandar PhD[1], Patrick Ryan, PhD[1,2], Chunhua Weng, PhD[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY USA;
[2]Janssen Research and Development, Titusville, NJ USA

Abstract

The Generalizability Index for Study Traits (GIST) has been proposed recently for assessing the population representativeness of a set of related clinical trials using eligibility features (e.g., age or BMI), one each time. However, GIST has not yet been evaluated. To bridge this knowledge gap, this paper reports a simulation-based validation study for GIST. Using the National Health and Nutrition Examination Survey (NHANES) data, we demonstrated the effectiveness of GIST at quantifying the population representativeness of a set of related trials that differ in disease domains, study phases, sponsor types, and study designs, respectively. We also showed that among seven example medical conditions, the GIST of age increases from Phase I trials to Phase III trials in the seven disease domains and is the lowest in asthma trials. We concluded that GIST correlates with simulation-based generalizability results and is a valid metric for quantifying population representativeness of related clinical trials.

Introduction

Randomized controlled trials (RCTs) have been widely regarded as the gold standard in medical research [1]. To ensure the internal validity of a clinical study when testing the efficacy of a treatment, clinical trialists often use restrictive eligibility criteria for participant selection [2]. However, unjustified exclusion criteria may unfairly deprive the opportunity of patients to benefit from the trial, and more importantly, compromise the generalizability of its results to the real-world patient population [3]. Consequently, many FDA-approved medications were later withdrawn from the market due to safety problems that had not been detected in pre-marketing clinical trials but only apparent after exposing the medications to a broader patient population [4, 5].

To assess the population representativeness of clinical trials, researchers may compare the study population of a single trial with a convenience sample of the real-world patient population [6, 7]. Most of the generalizability assessment studies identified in the literature focus on posteriori generalizability and thus can be conducted only after the conclusion and publication of a trial. In contract, priori generalizability, whose focus is on the eligibility of participants, can be assessed during trial design. Posteriori generalizability is almost always lower than priori generalizability because eligibility criteria always subsume the characteristics of study participants [8]. In addition, our previous study found that many trials, especially those on the same medical condition, often use similar or identical eligibility criteria, indicating that the generalizability issue may not be only at the individual trial level, but also across a whole body of trials for a clinical domain at the research community level [9].

To facilitate a priori generalizability assessment, a method was recently published for systematically assessing the population representativeness of a set of related trials. This method compares the aggregate target populations of all the trials under consideration, which characterize the patients who can be enrolled in these trials according to the inclusion and exclusion criteria, with the "real-world" patient population from electronic health records (EHRs) [10]. The Generalizability Index for Study Traits (GIST) was initially introduced along with this method [10]. This paper reports the initial study evaluating the validity and effectiveness of GIST.

Background

GIST is a mathematical function for quantifying the collective population representativeness of a set of clinical trials with reference to the real-world patient population measured by a single quantitative eligibility feature. There are three parameters in the GIST function: i.e., the real-world patient population (PP), the target population (TP) of a trial set, and an eligibility feature (not shown in the formula for brevity). The GIST metric is conceptually similar to Weisberg *et al.*'s model of patient selection bias in a trial using a counterfactual framework, which takes into account the proportion of patients with an adverse event incidence and the probability of such patients being selected by a trial [11]. GIST score ranges between 0 and 1, with 1 being most generalizable and 0 being least generalizable. It first discretizes the value range of an eligibility feature into consecutive non-overlapping value intervals and then sums the percentage of studies that recruit patients in each interval multiplied by the percentage of patients in the real-world population observed in that interval across all the intervals. GIST can quantify the representativeness of the target population (TP) for the patient population (PP) with respect to an eligibility feature such as age. The mathematical formula of GIST is:

$$\boldsymbol{GIST(TP,PP)} = \sum_{i=1}^{N} \frac{\sum_{j=1}^{T} \boldsymbol{I}([i_{low}, i_{high}] \subset w_j)}{T} * \frac{\sum_{k=1}^{P} \boldsymbol{I}(i_{low} \leq y_k < i_{high})}{P} \quad (1)$$

where N is the number of distinct value intervals of the quantitative eligibility feature under consideration, T is the number of trials in the trial set included for aggregate analysis, P is the number of patients in the patient population PP, w_j is the inclusion value interval of the quantitative feature for the j^{th} study, such that indicator I can be defined as j^{th} study interval subsumes the i^{th} interval's low and high boundary, and y_k is the observed value of the quantitative feature for the k^{th} patient such that an indicator I can be defined when k^{th} patient has a value of the quantitative feature falling within the i^{th} interval. Note that the GIST metric can also be applied to categorical variables, whereby the value intervals are integers.

Methods

A conceptual framework for validating GIST

Ideally, the GIST metric can be validated by taking two or more trial sets with known different generalizability with respect to an eligibility feature and assessing if the difference in their GIST scores correlates with the expected generalizability differences. As illustrated in **Figure 1(a)**, given TP1 and TP2 such that the population representativeness of TP1 is known to be better than TP2 with respect to a certain eligibility feature, GIST can be validated by verifying if GIST(TP1, PP) > GIST(TP2, PP). However, it is not feasible to obtain TP1 and TP2 because we have no evidence yet what kinds of trials have better generalizability with respect to a certain eligibility feature. Meanwhile, each trial set is affiliated with three different patient populations, i.e., the real-world patient population, the target population constructed from eligibility criteria descriptions, and the study population that includes all enrolled patients in the trial. Our method for validating the GIST metric is to simulate different patient populations that would result in a known generalizability difference with respect to the same trial set and assess if the difference of their GIST scores correlates with the expected difference. Therefore, we simulated a patient population that has better generalizability than the real-world patient population for the same trial set through weighted sampling of the real-world patient population. The relationships of these populations in our simulation-based validation method are illustrated in **Figure 1(b)** and their definitions are provided as follows:

Clinical trial target population (P0): the patients being sought as defined in the clinical trial eligibility criteria.

"Real-world" patient population (P1): the patients to whom the results of clinical trials are intended to be applied. We can only approximate its definition given available data resources about these patients.

Weighted "real-world" patient population (P2): the patients sampled from the "real-world" patient population based on the percentage of trials considering these patients for inclusion.

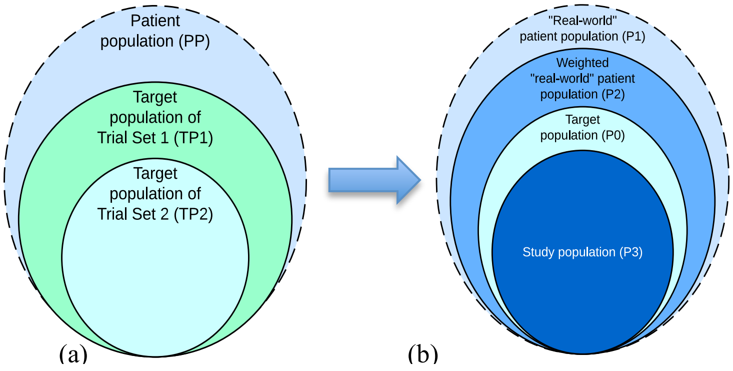
Patient population (PP)
Target population of Trial Set 1 (TP1)
Target population of Trial Set 2 (TP2)
(a)
"Real-world" patient population (P1)
Weighted "real-world" patient population (P2)
Target population (P0)
Study population (P3)
(b)

Figure 1. The conceptual framework for validating GIST.

Clinical trial study population (P3): the study participants who are actually enrolled in a clinical trial. Compared to P1 and P2, the study population P3 maximally (if not perfectly) reflects the target population (P0) because all enrolled study subjects should meet the eligibility criteria that define the target population.

For example, diabetes research may target Type 2 diabetes mellitus (T2DM) patients by defining P0 as "patients with HbA1c above 7.5%", while the real-world T2DM patients (P1) may be those patients whose HbA1c is above 7.0%, and the clinical trial study population (P3) may be a subset of real-world diabetes patients whose HbA1c is above 8.0%. Therefore, P1 subsumes P0, which further subsumes P3.

In [10], to reveal the population representativeness problem at a research community level, we aggregated multiple related clinical trials and used the distribution of trials over HbA1c values to represent the collective target population, i.e., the percentage of trials considering patients with a certain HbA1c value. In this work, we created P2 by sampling real-world patients based upon the percentage of trials considering them, thereby bringing the real-world patient population (P1) closer to the target population (P0). As such, P2 should be better represented in the collective target population (P0) than P1. Using P0 as the target, increasing generalizability will be observed in P1, P2, and P3, in order. Therefore, if the GIST scores for them follow GIST(P0,P1) < GIST(P0,P2) < GIST(P0,P3), we can conclude that GIST is a valid metric in quantifying the population representativeness of a trial set.

We first validated the GIST metric using this simulation-based method. Then we compared the GIST scores of trial sets that differ in their disease domains, sponsor types, study phases, and study designs. We hypothesized that:

Hypothesis #1: Weighted "real-world" patient population can serve as a good reference standard for validating GIST's suitability for indicating the population representativeness of a set of related clinical trials, one eligibility-feature each time.

Hypothesis #2: GIST correlates with the population representativeness of a set of related trials.

To profile the patient populations used for GIST evaluation, we used the population health data from the National Health and Nutrition Examination Survey (NHANES), a continuous cross-sectional health survey conducted by the National Center for Health Statistics of Centers for Disease Control and Prevention (CDC) [12]. NHANES evaluates a stratified multistage probability sample of the non-institutionalized population of the United States. The survey samples are first interviewed at home, followed by a physical and a laboratory test in a mobile examination center. Its rigorous quality control standards ensure national population representativeness and high-quality data collection.

Figure 2 shows the data collection and analysis pipeline employed in this study. We first extracted patient data from the NHANES database downloaded from the CDC website. We then retrieved the clinical trial summary text from ClinicalTrials.gov. We extracted and aggregated baseline characteristics of enrolled patients in T2DM trials with results. All the data were extracted with R and Python scripts, and subsequently stored in a MySQL database. After processing the data, we first evaluated the GIST metric. Then, we used GIST to compare population representativeness of trial sets of various characteristics. We will explicate each step as follows:

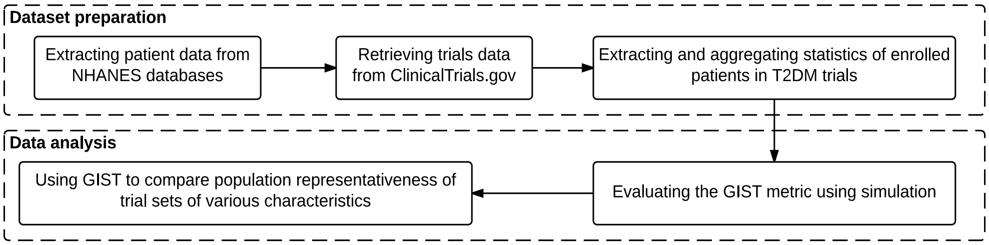
Dataset preparation
Extracting patient data from NHANES databases
Retrieving trials data from ClinicalTrials.gov
Extracting and aggregating statistics of enrolled patients in T2DM trials
Data analysis
Using GIST to compare population representativeness of trial sets of various characteristics
Evaluating the GIST metric using simulation

Figure 2. The data collection and analysis pipeline.

Step 1: Extracting patient data from NHANES databases

To ensure the statistical power of the analysis, we identified seven medical conditions in NHANES, each having more than 1,000 samples after combining data in multiple survey cycles. They were Type 2 diabetes mellitus (T2DM), asthma, arthritis, depression, sleep disorders, heart attack, and stroke. In the following, we will describe how we extracted and processed the NHANES data for these seven selected medical conditions.

T2DM: We combined the results of the Diabetes questionnaire of five continuous survey cycles between 2003 and 2012 and identified 3,304 diabetics who had their diabetes confirmed by a health professional and one HbA1c (Glycohemoglobin) measurement. As NHANES does not distinguish between two subtypes of diabetes, we employed a method used by Dodd *et al.* [13] to further identify 3,082 T2DM patients after excluding 222 samples with Type 1 diabetes who were (1) first diagnosed with diabetes before age 30; and (2) taking insulin. The rationale is that as one grows older, his/her lifestyle (e.g., dietary habits) will play a more important role in developing T2DM. Three quantitative eligibility features that are frequently used in T2DM trials, i.e., age (99.0%), HbA1c (53.6%), and Body Mass Index (BMI) (46.6%), were used for GIST evaluation. We combined the laboratory test results on HbA1c and examination data on body measures for five continuous survey cycles between 2003 and 2012. Out of the 3,082 T2DM samples, 2,695 had no missing values for age, HbA1c, and BMI. We used Chi-square test on two categorical variables, i.e., "gender" and "ethnicity", to test the representativeness of these 2,695 patients with no missing values for all the 3,082 patients. No statistically significant difference was found (*P-value* > 0.05). Therefore, we concluded that these 2,695 patients is a representative sample of all the T2DM patients in NHANES and included them in our further analysis.

Depression: NHANES started to conduct interviews on depression in the survey cycle of 2005-2006. We combined the results of the Depression Screener questionnaire of four continuous survey cycles between 2005 and 2012. The

Depression Screener questionnaire uses a 9-item screening instrument that asks questions about the frequency of symptoms of depression over the past 2 weeks. For example, participants were asked how frequently they "have little interest in doing things" or "feel tired or have little energy." Every question was rated from "0" to "3", where "0" means "not at all" and "3" means "nearly every day." Employing a method used by Xiao *et al.* [14], we identified 1,884 depressive participants who have a combined score of 10 or higher for the nine questions.

Sleep disorders: NHANES started to conduct interviews on sleep disorders in the 2005-2006 survey cycle. We combined the results of the Sleep Disorders questionnaire of four continuous survey cycles between 2005 and 2012 and identified 1,816 participants who were told by a doctor or a health professional to have sleep disorders.

Asthma, **arthritis**, **heart attack**, and **stroke**: After combing the results of Medical Conditions questionnaire of five continuous survey cycles between 2003 and 2012, we identified 7,009, 7,449, 1,235, 1,119 participants who were told by a doctor or health professional to have asthma, arthritis, heart attack, and stroke, respectively. It is worth noting that NHANES does not provide other laboratory tests to further validate these conditions.

To account for oversampling, non-response, and post-stratification, NHANES assigned each participant a two-year sample weight (WTMEC2YR), which is the number of people in the U.S. national population that each participant can represent. According to the analytical guideline of NHANES [15], we calculated eight-year sample weight WTMEC8YR (1/4 * WTMEC2YR) for depression and sleep disorders patients, because their data in four survey cycles were combined. For the other five conditions, we calculated ten-year sample weight WTMEC10YR (1/5 * WTMEC2YR), because data in five survey cycles were combined for them. After applying the normalized sample weights in the analysis, the patients in NHANES can represent the U.S. non-institutionalized population in the midpoint of the combined survey period.

Step 2: Retrieving trials data from ClinicalTrials.gov

To facilitate large-scale systematic analysis of population representativeness of related clinical trials, we have built a computable repository of clinical trials called COMPACT, which stores fine-grained eligibility features and descriptive characteristics of all the trials in ClinicalTrials.gov [16]. COMPACT indexed trials by medical conditions, allowing efficient aggregate analysis of trials on the same condition. Based on COMPACT, we have built a Web-based visual analytic tool of eligibility features in clinical trials called VITTA [17]. VITTA allows its users to select trials of a particular medical condition, refine the selection of trials by various characteristics, and profile the collective target population with a single eligibility feature.

For each medical condition, from COMPACT we retrieved interventional studies that had their start date falling in the survey years of NHANES. Corresponding patient data were obtained. For example, because patient data with sleep disorders were obtained from NHANES between 2005 and 2012, we also retrieved interventional studies on sleep disorders with the start date between January 2005 and December 2012 from the COMPACT database.

Step 3: Retrieving a convenience sample of enrolled patients in T2DM trials

As the study population of enrolled patients should well represent the target population of a trial, the collective study population should yield better population representativeness than the "real-world" patient population. To test whether GIST score can reflect this expected difference, we retrieved the results data of T2DM trials between 2003 and 2012 that reported summary data of their enrolled patients in ClinicalTrials.gov. The summary data must report the number of participants, mean and standard deviation (SD) value of at least one of age, HbA1c, and BMI to be included. We aggregated the mean and SD of age, HbA1c, and BMI separately using the following formula (adapted from [18]), where *T* is the number of studies,

$$Weighted\_mean = \frac{\sum_{i=1}^{T}(mean_i * number\_participants_i)}{\sum_{i=1}^{T} number\_participants_i} \quad (2)$$

$$Weighted\_SD = \sqrt{\frac{\sum_{i=1}^{T}(SD_i^2 * (number\_participants_i - 1))}{\sum_{i=1}^{T}(number\_participants_i - 1)}} \quad (3)$$

Table 1 shows the number of T2DM trials that reported summary data of age, HbA1c, and BMI for their enrolled patients, the total enrollments of these trials, and aggregated mean and SD values of age, HbA1c, and BMI, respectively. Given that only a small number of trials reported summary data of their enrolled patients in ClinicalTrials.gov, these aggregated data represent a convenience sample of enrolled patients in all the T2DM trials.

Table 1. The mean and SD of age, HbA1c, and BMI of the convenience sample of enrolled patients in T2DM trials.

Variable	Number of trials	Number of patients	Mean	SD
Age	389	198,050	58.3	9.4
HbA1c	137	62,931	8.2	1.0
BMI	108	70,678	30.5	5.2

Step 4: Evaluating GIST using simulation

From COMPACT, we retrieved 2,731 interventional studies on T2DM with a start date falling between 01/2003 and 12/2012. The number of T2DM trials specifying permissible values for age, HbA1c, and BMI was 2,702 (99.0%), 1,463 (53.6%), and 1,274 (46.6%), respectively. We formed one trial set for each of the three features and included them for generating the distribution of trials (P0) over age, HbA1c, and BMI, respectively. These distributions were used for evaluating the GIST metric. For each feature, we generated three patient samples as follows:

"Real-world" patient population (P1): A random sample of 10,000 patients from the T2DM patients in NHANES using normalized NHANES sample weight with replacement. NHANES sample weight is the number of patients in the U.S. national population that one survey participant can represent. Therefore, this sample can represent the "real-world" T2DM patients.

Weighted "real-world" patient population (P2): We generated a random sample of 10,000 patients from T2DM patients in NHANES considering both NHANES sample weight and the percentage of trials that consider such patients. Specifically, we separately normalized NHANES sample weight (i.e., WTMEC10YR) and percentage of trials (P0), and then used the average of these two normalized weights to sample the T2DM patients in NHANES. As such, the sample of weighted "real-world" patients generated by oversampling "real-world" patients who are considered by more trials and under-sampling "real-world" patients who are considered by fewer trials would be better represented in the target population of the trials (P0) than the "real-world" patient population (P1). Note that the values in both the normalized sample weights and normalized percentages of trials added up to 100%, while the sum of values in P0 did not.

Study population of clinical trials (P3): A random sample of 10,000 enrolled patients generated using Gaussian distribution with the mean and SD of the convenience sample of enrolled patients in T2DM trials (from Step 3).

In the ideal situation, the study population should represent the target population, assuming the enrolled patients match the eligibility criteria perfectly. Therefore, P3 should have the best generalizability of the target population among P1, P2 and P3. The same trial set should have better population representativeness for P2 than P1. For each feature (i.e., age, HbA1c, and BMI), we calculated the GIST scores for three patient samples. Taking sampling variability into account, we ran the experiment for 100 times. Note that P0 for each feature remained the same in all the experiments, whereas P1, P2, and P3 were generated once for each feature in an experiment. If the calculated GIST scores in all the experiments consistently follow GIST(P0,P1) < GIST(P0,P2) < GIST(P0,P3), GIST correlates the expected difference of three patient samples and is therefore a valid metric for assessing the population representativeness of a given patient population in a given trial set.

Step 5: Comparing population representativeness of trial sets of various characteristics

To reveal the population representativeness problem at the research community level, we calculated the GIST score of age for each of the seven previously selected conditions: T2DM, depression, asthma, sleep disorders, arthritis, heart attack, and stroke. To compare population representativeness of different types of trials, we further performed stratification analysis on study phases, sponsor types, and study designs across multiple conditions. We used the GIST scores of age for trial sets that differ in these trial characteristics to compare their population representativeness.

Results

Evaluation results of GIST using simulation

For each of the three eligibility features (i.e., age, HbA1c, and BMI), we generated three patient samples (i.e., P1, P2, and P3, defined in Step 4 of the Methods Section) that have known differences in generalizability for collective target population (P0) of the same set of trials in one experiment and ran the same experiment for 100 times. Even though the difference between P1 and P2 may be minor if most trials accept broad range of values, the GIST metric should still capture the difference. To illustrate the differences among three patient samples, we visualized in **Figure**

3 the distribution of P1, P2, and P3 against the target population of T2DM trials (P0) for age, HbA1c, and BMI in the pilot experiment. The widths of value intervals for age, HbA1c, and BMI are 1, 0.5, and 1, respectively. In each sub-figure, the green dot-and-dashed curve represents the target population of T2DM trials, i.e., the percentage of trials allowing a value interval. The blue solid curve represents the distribution of patients in the sample of "real-world" patients (P1) over consecutive non-overlapping value intervals of a feature. The red dotted solid curve represents the distribution of patients in the weighted sample of "real-world" patients (P2). The light blue solid curve with big dot represents the distribution of patients in the sample of enrolled patients in T2DM trials (P3). There are two y-axes: the left one is for the three sample patient populations (i.e., P1, P2, and P3) and the right one is for the target population of clinical trials (P0).

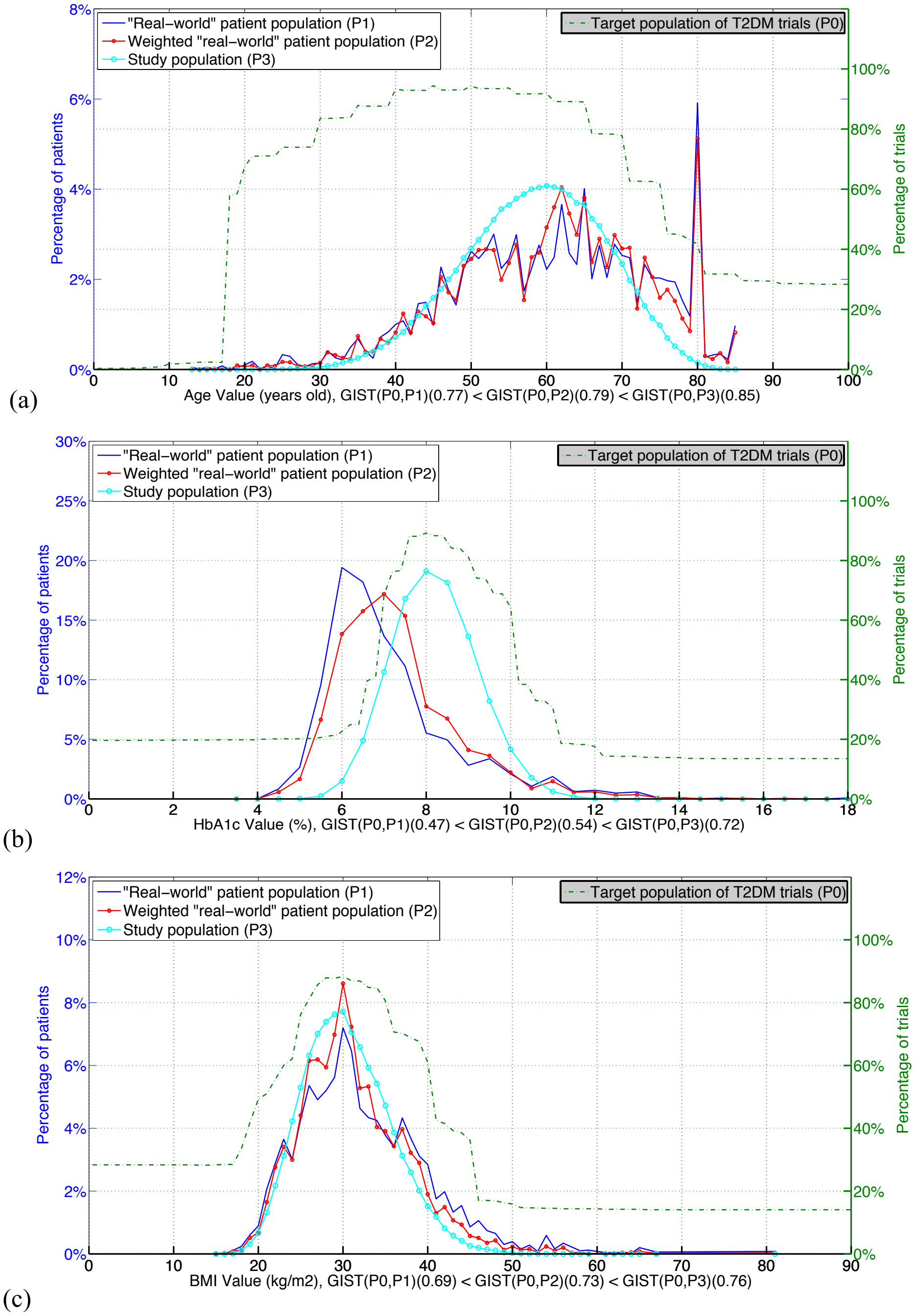

Figure 3. Visualization of three patient populations (P1, P2, and P3) and the target population of T2DM trials (P0) with respect to (a) age, (b) HbA1c, and (c) BMI, respectively.

As can be seen in **Figure 3(a)**, both distributions of P1 and P2 peak at age 80, which is considered by only about 40% of trials. The distributions of P1 and P2 are similar in the visualization, but a statistically significant difference between them was observed in two-sample Kolmogorov-Smirnov test (test statistic = 0.028 > 0.014 = critical value, P-value < 0.05). **Figure 3(b)** shows the visualization of three patient samples and collective target population regarding HbA1c. The distribution of P3 aligns with P0 better than that of P1 and P2. The peak of the distribution of P2 stands in between that of P1 and P3, confirming that the same set of trials does have better generalizability for P2 than P1. **Figure 3(c)** shows the visualization for BMI. We can see that all three distributions of P1, P2, and P3 peak at about 30 kg/m^2, where the peak of P2 is higher than that of P1. As BMI value increases from 30 kg/m^2, the curve of P2 gradually drops below that of P1. The reason is that the weight (i.e., target population) we used to generate P2 decreases from 30 kg/m^2 onwards. In this pilot experiment, the GIST scores of all the three features follow GIST(P0,P1) < GIST(P0,P2) < GIST(P0,P3).

To ensure stability of our evaluation, we ran the aforementioned experiment for 100 times for each eligibility feature. **Table 2** shows the mean and SD of 100 GIST scores of P1, P2, and P3 for three quantitative eligibility features. SD values are very low (between 0.0002 and 0.0029), indicating that the GIST scores for all three features consistently followed GIST(P0,P1) < GIST(P0,P2) < GIST(P0,P3) in these experiments. The results were consistent with the expected differences in the population representativeness for three patient samples. Thus, we demonstrated the validity of GIST at assessing the population representativeness of related clinical trials. We also confirmed our *Hypothesis #1* that weighted "real-world" patient population could serve as a good reference standard for validating GIST's suitability for indicating the population representativeness of a set of related trials, one eligibility-feature at a time.

Table 2. Mean and SD of GIST(P0,P1), GIST(P0,P2), and GIST(P0,P3) in 100 experiments for each eligibility feature ranked in ascending order of mean GIST scores.

Eligibility Features	**Mean of GIST(P0,P1)**	**SD of GIST(P0,P1)**	**Mean of GIST(P0,P2)**	**SD of GIST(P0,P2)**	**Mean of GIST(P0,P3)**	**SD of GIST(P0,P3)**
HbA1c	0.47	0.0029	0.55	0.0025	0.72	0.0007
BMI	0.69	0.0020	0.73	0.0019	0.76	0.0005
Age	0.77	0.0021	0.79	0.0018	0.85	0.0002

GIST scores of trial sets of various characteristics

To uncover the population representativeness issue at the research community level, we used GIST to compare the population representativeness with respect to age for seven previously chosen medical conditions. **Table 3** shows the GIST scores of age for these seven medical conditions ranked in ascending order (Column 5). The numbers of patients in the U.S. national population (Column 3) are the sum of the normalized sample weights (i.e., WTMEC10YR or WTMEC8YR) of the survey participants with the corresponding condition. The GIST score of age can assess how the target population of a set of trials using age as an eligibility criterion (Column 4) represents the "real-world" patient population (Column 3). We observed that among all the conditions, asthma trials had the worst population representativeness (0.54), while heart attack trials had the best population representativeness (0.89). The GIST of age for trials on the other five conditions ranged between 0.74 and 0.83, showing relatively good population representativeness.

Table 3. The GIST score of age for seven medical conditions ranked in ascending order.

Medical condition	**Number of samples in NHANES (year range)**	**# of patients in the U.S. national population**	**# of trials with age (year range)**	**GIST of age**
Asthma	7,009 (2003-2012)	42,035,521	1,459 (2003-2012)	0.54
Sleep disorders	1,816 (2005-2012)	17,517,187	995 (2005-2012)	0.74
Depression	1,884 (2005-2012)	15,366,622	1,956 (2005-2012)	0.75
T2DM	2,695 (2003-2012)	15,575,484	2,702 (2003-2012)	0.77
Arthritis	7,449 (2003-2012)	52,355,612	2,190 (2003-2012)	0.82
Stroke	1,119 (2003-2012)	6,174,893	1,098 (2003-2012)	0.83
Heart attack	1,235 (2003-2012)	7,387,760	638 (2003-2012)	0.89

Roumiantseva *et al.* have found that industry-sponsored studies differ systematically from government-sponsored studies in study type, interventions, and condition studied [19]. We are also interested in exploring the difference of population representativeness for trial sets of various characteristics.

Table 4 gives the GIST scores of age for trial sets in different study phases, study sponsors, and study designs on seven medical conditions, horizontally ordered in the same order as Table 3. For all the conditions, the GIST score of age increases from Phase I to Phase III. This is in accordance with the fact that Phase I trials aim to establish initial safety and efficacy profile of a treatment in a small group of patients, while Phase III trials seek to test the treatment with a large groups of people to confirm its safety and efficacy.

According to the GIST scores, industry-sponsored trials have better population representativeness than NIH-sponsored trials for asthma, sleep disorders, depression, T2DM, and arthritis. In general, randomized trials have a slightly better population representativeness than non-randomized trials except for asthma trials. With regards to primary purpose (study design), treatment and diagnostic trials have better population representativeness than prevention and basic science trials. These results confirmed our *Hypothesis #2* that GIST correlates with the population representativeness of a set of related trials.

Table 4. The GIST scores of age for trials in different phases, sponsors, and study designs on seven medical conditions. The number of trials is enclosed by parentheses.

Trial Characteristics	**Asthma**	**Sleep disorders**	**Depression**	**T2DM**	**Arthritis**	**Stroke**	**Heart attack**
	GIST score of age (# of trials)						
Study Phase							
Phase I	0.53 (179)	0.68 (76)	0.67 (229)	0.60 (368)	0.71 (261)	0.75 (183)	0.79 (60)
Phase II	0.58 (410)	0.75 (188)	0.72 (408)	0.77 (517)	0.84 (566)	0.84 (319)	0.87 (172)
Phase III	0.59 (336)	0.80 (193)	0.81 (384)	0.87 (766)	0.86 (582)	0.85 (212)	0.92 (160)
Phase IV	0.52 (231)	0.69 (136)	0.78 (318)	0.80 (484)	0.83 (404)	0.82 (131)	0.93 (173)
Sponsor Type							
NIH	0.31 (27)	0.64 (10)	0.65 (44)	0.57 (35)	0.81 (29)	0.86 (27)	1.00 (2)
Industry	0.60 (778)	0.80 (287)	0.80 (431)	0.80 (1464)	0.85 (1237)	0.85 (234)	0.92 (145)
U.S. Fed	0.45 (6)	0.75 (44)	0.93 (49)	0.81 (22)	0.89 (30)	0.93 (44)	1.00 (2)
Other	0.48 (648)	0.72 (654)	0.73 (1432)	0.73 (1181)	0.78 (894)	0.81 (793)	0.89 (489)
Study Design - Allocation							
Randomized	0.54 (1217)	0.75 (772)	0.76 (1525)	0.77 (2319)	0.83 (1661)	0.84 (867)	0.90 (554)
Non-Randomized	0.57 (140)	0.70 (93)	0.71 (201)	0.74 (217)	0.80 (295)	0.79 (104)	0.84 (43)
Study Design – Primary Purpose							
Treatment	0.56 (1077)	0.75 (767)	0.77 (1485)	0.80 (2043)	0.83 (1872)	0.83 (805)	0.89 (484)
Prevention	0.34 (81)	0.73 (24)	0.54 (139)	0.63 (217)	0.84 (62)	0.80 (151)	0.91 (65)
Diagnostic	0.56 (57)	0.83 (48)	0.70 (36)	0.77 (40)	0.79 (34)	0.79 (33)	0.95 (42)
Basic Science	0.53 (78)	0.57 (50)	0.59 (56)	0.64 (162)	0.66 (56)	0.57 (13)	0.54 (10)

Discussion

This study validated the effectiveness of GIST at assessing the population representativeness of related clinical trials. The GIST scores consistently agreed with the expected differences of population representativeness for three population samples across 100 experiments for all three selected eligibility features. We further demonstrated the effectiveness of GIST in comparing population representativeness of trial sets that differ in their characteristics such as disease domains, sponsor types, study phases, and study designs. Among seven medical conditions, asthma trials had the lowest GIST score of age, reflecting the concern in the respiratory medicine research community [20]. Meanwhile, the GIST metric was further validated by the increasing GIST scores of age from Phase I to Phase III across all the seven conditions. Note that GIST can also assess the population representativeness of a single clinical study, as its primary use case of a pharmaceutical company will be to assess the generalizability of a study it is currently designing or even a study that it already has completed.

In this work, we used NHANES to profile the "real-world" patient populations (P1) and weighted "real-world" patient populations (P2). Compared with EHR data, NHANES has several advantages. First, its sophisticated sampling mechanism ensures the population representativeness at the national level. In contrast, EHR data contain mostly diseased patients or patients receiving care and hence may be biased towards certain population subgroups. Second, structured survey data are readily analyzable, whereas EHR data often require preprocessing to address data quality problems. Therefore, NHANES is more cost-effective than EHR data for lightweight population-based studies. However, due to the limited data in NHANES, it may not be suitable for longitudinal analysis or studies on medical conditions that are not included in the interview questions.

Limitations

This study has limitations. We only included seven medical conditions that have a fair amount of patients (over 1,000) in NHANES. Ideally, more conditions should be analyzed. The self-reported medical conditions in NHANES may have resulted in some misclassification of samples. The GIST metric has intrinsic limitations. First, GIST does not take into account the enrollment value of a study. Currently, ClinicalTrials.gov has only one field for enrollment, which can be planned or actual. Quite a number of trials have not updated the planned enrollment with actual enrollment even after completion. Second, it does not consider the geographic location of the trial, which is one major factor for patient recruitment. Nevertheless, by aggregating many patients and clinical trials, we have minimized the impact of these factors for generating meaningful results in the research community level. Third, the GIST metric cannot reveal the reason behind the population representativeness problem. Visualization such as Figure 3 can serve as a good complement to GIST for assessing the population representativeness of related trials.

The long-term goals of this line of research

As the main purpose of most RCTs is to test the efficacy and safety of a treatment for a certain medical condition in people, it is often required to minimize confounding factors that may potentially affect the results. Therefore, it is a common practice that RCTs usually recruit patients who do not have comorbidities and are not too old or too sick to treat. Instead of enforcing a trial to be generalizable to the broad patient population, the goals of this line of research are (1) to improve the transparency of clinical trial eligibility criteria design biases across multiple studies; (2) to facilitate evidence-based data-driven precision design of clinical trial eligibility criteria [21]; and (3) to address the rising need for patient-centered outcomes research in the clinical trial domain. This information can be provided to clinical trial designers to help them better justify the trade-offs between the internal validity and the external validity when designing a new trial. This information can also help clinical investigators and policy makers efficiently identify population representativeness issues in clinical studies of certain characteristics and take measures accordingly. When applying clinical trial eligibility criteria to observational data, one can compare and contrast the effects observed between eligible and ineligible patients, which may reveal more profound problems in trial design and clinical research in general.

Future work

In the future, we plan to use GIST to identify restrictive features among multiple frequently used eligibility features among clinical trials of a certain medical condition. We will first leverage controlled terminologies such as SNOMED CT to develop structural, semantic, and lexical methods for meaningful aggregation of similar qualitative features. Then we will compute GIST scores to identify stringent features (with relatively low GIST scores). In this paper, we compared the population representativeness of trials on seven medical conditions with respect to age using GIST. However, this method is not efficient when more eligibility features are included in the analysis. Moreover, eligibility features may have inherent correlations. For example, a previous study has reported that impaired fasting glucose generally increases with age for diabetic patients [22]. Meanwhile, eligibility criteria may be operationalized as an interaction of multiple features, e.g., “pregnant female over 40 years old.” In the future, we will investigate how to assess the collective population representativeness using multiple eligibility features simultaneously.

Conclusions

In this work, we used real-world population-level data to validate a novel metric for quantifying the population representativeness of clinical trials. The study results confirmed that the GIST metric is reliable for its purpose. These findings suggested the future potential of a systematic approach for providing prognostic tools to facilitate the clinical trial design process as well as post hoc evaluations to investigate the generalizability of studies already underway or completed. By integrating the real-world experience of patients with the study design attributes of existing clinical trials, researchers designing new clinical studies can improve both the efficiency and generalizability of their designs with this proactive data-driven approach.

Acknowledgments

This study was sponsored by the U.S. National Library of Medicine grant **R01LM009886** (PI: Weng) and U.S. National Center for Advancing Translational Science grant **UL1 TR000040** (PI: Ginsberg).

References

1. From the NIH Director: The Importance of Clinical Trials [April 9, 2014]. Available from: http://www.nlm.nih.gov/medlineplus/magazine/issues/summer11/articles/summer11pg2-3.html.
2. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?". Lancet. 2005;365(9453):82-93.
3. Schmidt AF, Groenwold RHH, van Delden JJM, van der Does Y, Klungel OH, Roes KCB, et al. Justification of exclusion criteria was underreported in a review of cardiovascular trials. Journal of Clinical Epidemiology. 2014;67(6):635-44.
4. Wikipedia. List of withdrawn drugs [February 10, 2015]. Available from: http://en.wikipedia.org/wiki/List_of_withdrawn_drugs.
5. Wysowski DK, Swartz L. Adverse drug event surveillance and drug withdrawals in the United States, 1969-2002: the importance of reporting suspected reactions. Arch Intern Med. 2005;165(12):1363-9.
6. 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(10):627-30.
7. van de Water W, Kiderlen M, Bastiaannet E, Siesling S, Westendorp RG, van de Velde CJ, et al. External validity of a trial comprised of elderly patients with hormone receptor-positive breast cancer. J Natl Cancer Inst. 2014;106(4):dju051.
8. Blanco C, Olfson M, Goodwin RD, Ogburn E, Liebowitz MR, Nunes EV, et al. Generalizability of clinical trial results for major depression to community samples: results from the National Epidemiologic Survey on Alcohol and Related Conditions. J Clin Psychiatry. 2008;69(8):1276-80.
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. Weng C, Li Y, Ryan P, Zhang Y, Gao J, Liu F, et al. A Distribution-based Method for Assessing The Differences between Clinical Trial Target Populations and Patient Populations in Electronic Health Records. Applied Clinical Informatics. 2014;5(2):463-79.
11. Weisberg HI, Hayden VC, Pontes VP. Selection criteria and generalizability within the counterfactual framework: explaining the paradox of antidepressant-induced suicidality? Clin Trials. 2009;6(2):109-18.
12. CDC. National Health and Nutrition Examination Survey: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention;. Available from: http://www.cdc.gov/nchs/nhanes.htm.
13. Dodd AH, Colby MS, Boye KS, Fahlman C, Kim S, Briefel RR. Treatment approach and HbA1c control among US adults with type 2 diabetes: NHANES 1999-2004. Curr Med Res Opin. 2009;25(7):1605-13.
14. Xiao Y, Zhao N. Current cigarette use in rheumatoid arthritis patients: associated factors and a limited mediating role of depression. Rheumatol Int. 2015;35(7):1219-24.
15. CDC. National Health and Nutrition Examination Survey: Analytic Guidelines, 1999–2010 [September 2014]. Available from: http://www.cdc.gov/nchs/data/series/sr_02/sr02_161.pdf.
16. He Z, Carini S, Hao T, Sim I, Weng C. A Method for Analyzing Commonalities in Clinical Trial Target Populations. AMIA Annu Symp Proc. 2014:1777-86.
17. He Z, Carini S, Sim I, Weng C. Visual aggregate analysis of eligibility features of clinical trials. J Biomed Inform. 2015;54:241-55.
18. Pooled Variance [October 19, 2014]. Available from: http://en.wikipedia.org/wiki/Pooled_variance.
19. Roumiantseva D, Carini S, Sim I, Wagner TH. Sponsorship and design characteristics of trials registered in ClinicalTrials.gov. Contemp Clin Trials. 2013;34(2):348-55.
20. Herland K, Akselsen JP, Skjonsberg OH, Bjermer L. How representative are clinical study patients with asthma or COPD for a larger "real life" population of patients with obstructive lung disease? Respir Med. 2005;99(1):11-9.
21. Weng C, Li Y, Bigger JT, Zhang Y, Liu F, Gao J, et al. Using Electronic Data To Profile Population Health and Identify Clinical Evidence Gaps: Towards Precision Clinical Research Design. EDM Forum Stakeholder Symposium; June 2014; San Diego, CA.
22. Cowie CC, Rust KF, Byrd-Holt DD, Eberhardt MS, Flegal KM, Engelgau MM, et al. Prevalence of diabetes and impaired fasting glucose in adults in the U.S. population: National Health And Nutrition Examination Survey 1999-2002. Diabetes Care. 2006;29(6):1263-8.