# Comparing and Contrasting A Priori and A Posteriori Generalizability Assessment of Clinical Trials on Type 2 Diabetes Mellitus

**Zhe He, PhD[1], Arturo Gonzalez-Izquierdo, PhD[2], Spiros Denaxas, PhD[2], Andrei Sura, BS[3], Yi Guo, PhD[3], William R. Hogan, MD[3], Elizabeth Shenkman, PhD[3], Jiang Bian, PhD[3]**
**[1]Florida State University, Tallahassee, FL, USA; [2]University of College London, London, UK; [3]University of Florida, Gainesville, FL, USA**

**Abstract**

*Clinical trials are indispensable tools for evidence-based medicine. However, they are often criticized for poor generalizability. Traditional trial generalizability assessment can only be done after the trial results are published, which compares the enrolled patients with a convenience sample of real-world patients. However, the proliferation of electronic data in clinical trial registries and clinical data warehouses offer a great opportunity to assess the generalizability during the design phase of a new trial. In this work, we compared and contrasted a priori (based on eligibility criteria) and a posteriori (based on enrolled patients) generalizability of Type 2 diabetes clinical trials. Further, we showed that comparing the study population selected by the clinical trial eligibility criteria to the real-world patient population is a good indicator of the generalizability of trials. Our findings demonstrate that the a priori generalizability of a trial is comparable to its a posteriori generalizability in identifying restrictive quantitative eligibility criteria.*

**Introduction**

Clinical trials, which test the efficacy and safety of an intervention (e.g., medication, device, procedure, and behavioral change), are indispensable tools for evidence-based medicine [1]. However, the generalizability of clinical research studies has long been a concern [2]. For example, elderly patients are reported to be underrepresented in clinical trials across major medical conditions, including cardiovascular diseases [3], cancers [4], dementia [5], and diabetes [6]. Most research on the generalizability of clinical studies has focused on the *a posteriori* generalizability (i.e., the representativeness of enrolled participants), which compares the characteristics of enrolled patients of a study with a convenience sample of a real-world patient population [7] or those in other studies [8]. For example, van der Water *et al*. evaluated the external validity of a cancer clinical trial by comparing the socioeconomic status, number of comorbidities, treatments, and various stage information between the enrolled patients and the patients in the Netherlands Cancer Registry [7]. Cahan and colleagues proposed an *a posteriori* generalizability score that incorporates demographic information, clinical attributes, and clinical settings to compare a trial to multiple target clinical scenarios in other trials [8]. However, with such an *a posteriori* approach, the generalizability issue is not detected before the conclusion of the studies. Typical clinical trials cost over hundreds of millions dollars or more and take 10 - 17 years to complete [9]. Thus, it is crucial to assess a clinical trial's generalizability before conducting the trial. Nevertheless, existing methods for assessing the *a priori* generalizability (i.e., the representativeness of eligible participants) have historically been scarce and laborious [10].

The rapidly growing amount of electronic patient data such as electronic health records (EHR) data, presents an unprecedented opportunity for optimizing eligibility criteria in the design phase of a new trial towards balanced internal and external validity (i.e. generalizability) [11]. In recent years, a suite of informatics methods has been introduced to quantify the population representativeness of clinical studies and characterize underrepresented population subgroups [12-14]. Notably, the Generalizability Index on Study Traits (GIST) metric quantifies the *a priori* generalizability of clinical trials with respect to selected quantitative eligibility criteria that specify a permissible value range (e.g., HbA1c > 7%), one at a time [12]. The extension of GIST, mGIST [15], can quantify the population representativeness of clinical trials with joint use of multiple criteria of interest. Both GIST and mGIST focus on the generalizability assessment at the disease domain level (i.e., assessing the generalizability of trials targeting the same disease). Recently, Sen and colleagues introduced GIST 2.0 as a scalable multivariate metric for quantifying the population representativeness of individual clinical studies by explicitly modeling the dependencies among all eligibility criteria [16]. However, to the best of our knowledge, no work has compared the *a priori* and the *a posteriori* generalizability assessment of trials.

Diabetes, recognized as an important public health problem by the World Health Organization [17], has caused 1.5 million deaths in 2012 alone and may, over time, lead to serious damage to the heart, blood vessels, eyes, kidneys,

and nerves. More than 400 million people live with diabetes [18]. Type 2 diabetes mellitus (T2DM), which can be developed at any age, accounts for 90% - 95% of people who have diabetes [19]. Many countries, including United States (US) and United Kingdom (UK), have invested heavily in research on treating and controlling diabetes [20, 21]. In this study, we compared and contrasted *a priori* (i.e., using GIST, Weng et al. [12]) and *a posteriori* (i.e., using van de Water et al. [7]) generalizability of T2DM clinical trials. We hypothesize that the *a priori* generalizability of a trial is comparable to its *a posteriori* generalizability in identifying certain restrictive quantitative eligibility criteria. To enable such a comparison, we will use univariate GIST metric to assess *a priori* generalizability of T2DM trials with respect to three most frequently used quantitative eligibility criteria, age, HbA1c, and BMI. We chose GIST rather than its multivariable extensions, mGIST, as we want to compare individual variables' *a priori* generalizability with their *a posterior* generalizability independently. We used the eligibility criteria of T2DM trials registered on ClinicalTrials.gov to profile the study populations, and extracted the published summary-level statistics of the enrolled patients. We compared the *a priori* (based on the study populations) and the *a posteriori* (based on the enrolled patients) generalizability of US-based T2DM trials using the target population profiled by the T2DM patients in the OneFlorida Data Trust [22], and further validated the results using UK-based T2DM trials with the target population profiled by the patients from the CALIBER research platform [23]. The OneFlorida Clinical Research Consortium (CRC) is one of the 13 Clinical Data Research Networks (CDRN) in the United States funded by the Patient-Centered Outcome Research Institute (PCORI) as part of the National Patient-Centered Clinical Research Network (PCORnet). The CALIBER resource generates and investigates deep, longitudinal phenotypic data from linked electronic health records for people registered in participating clinical practices in UK. The four main data sources include primary care EHR, hospital billing data, and death certificate records. Our work will inform the research community of the difference of *a priori* and *a posteriori* generalizability of T2DM trials with respect to quantitative eligibility features.

## Background

### *ClinicalTrials.gov and the COMPACT Database*

ClinicalTrials.gov, created and maintained by the National Library of Medicine, is a clinical study registry in the United States. Since September 2007, all the United States-based clinical trials of FDA-regulated drugs, devices, and biologic products must be registered in ClinicalTrials.gov prior to participant recruitment. In September 2016, the United States Department of Health and Human Services issued the final rule of that expands the regulatory procedure for trial registration and summary results reporting [24]. Mandated by the final rule, trial sponsors are required to report summary statistics on race, ethnicity, and other measures assessed at baseline that are used in analyzing a primary outcome measure on ClinicalTrials.gov [24]. As ClinicalTrials.gov is the largest clinical trial registry in the world with a long history of system operation and management [25], many international trials are also registered in ClinicalTrials.gov. As of February 6, 2017, as many as 236,212 studies with locations in all 50 states in the US as well as in 195 countries have been registered. Study summaries are semi-structured in ClinicalTrials.gov: study descriptors such as study phase (i.e., Phase I, II, III, and IV), intervention type (e.g., drug, device, biologic product), locations, are stored in structured fields, whereas eligibility criteria are largely free-text.

To facilitate *a priori* generalizability analysis, we leveraged the numeric expression extraction tool "Valx" [26, 27] and the frequent tag mining tool [28] to transform study summaries in ClinicalTrials.gov into a relational database, and we named it "COMPACT" [29]. COMPACT contains various study descriptors (e.g., study phase, intervention), and numeric eligibility features. With Valx, different names for the same variable in the eligibility criteria, such as "hemoglobin A1c", "HbA1c", "Glycohemoglobin", are unified. Different measurement units are also unified. COMPACT indexes trials by medical conditions.

### *OneFlorida Clinical Research Consortium and Data Trust*

The OneFlorida CRC is a collaborative statewide network that seeks to improve health research capacity and opportunities in the State of Florida through the facilitation of clinical and translational research in communities and health care settings. OneFlorida includes nine unique health systems that provides care for ~9.7M or 48% of all Floridians through 4,100 physician providers, 1,240 clinic/practice settings and 22 hospitals with a catchment area covering all 67 Florida counties. In 2015, OneFlorida became one of 13 PCORI-funded clinical data research networks in the US.

The OneFlorida Data Trust is the centerpiece of the OneFlorida CRC and is the informatics infrastructure that supports pragmatic trials; comparative effectiveness research, implementation science, and other research in OneFlorida. The OneFlorida Data Trust currently contains collated EHR, health care claims, and other data on a

broad-based, unselected population of ~10 million people in Florida. The data are limited to a Health Insurance Portability and Accountability Act (HIPAA) Limited Data Set (LDS), which restricts the types of protected health information (PHI) to only dates (e.g., birthdates and dates of service) and location (to the zip code level).

***CALIBER (Clinical research using LInked Bespoke studies and Electronic health Records)***

CALIBER is a unique research platform consisting of 'research ready' variables extracted from linked EHR from primary and secondary care, social deprivation information and cause-specific mortality data in UK. Led from the University College London Institute of Health Informatics and the Farr Institute of Health Informatics Research, London, CALIBER enables researchers to recreate the longitudinal journey of patients through care pathways to study disease onset and progression. This research platform accesses linked electronic health records and recreates the healthcare pathways of approximately 10 million patients with 400 million person years of follow up. The aim of CALIBER is to foster an open community developing methods and tools to accelerate replicable science across all clinical and scientific disciplines spanning the translational cycle (from drug discovery through to public health). The resource consists of disease and risk factor phenotyping algorithms, methods [30], tools and scripts, specialized infrastructure and training and support. All finalized EHR phenotyping algorithms are provided in an open-access Portal (https://www.caliberresearch.org/portal) for researchers to extent and re-use.

**Methods**

We first define the three patient populations for a clinical trial:

- ***Target population****:* patients to whom the results of the clinical study are intended to apply. The target population can only be approximated with available patient data.
- ***Study population****:* patients who are eligible for the study based on the study inclusion and exclusion criteria.
- ***Enrolled patients****:* patients who are enrolled in the clinical study. It is a subset of the study population.

Figure 1 illustrates the analytical workflow of this study. We first retrieved interventional clinical studies on T2DM in the US and UK from the COMPACT database. In this work, we chose to focus on three major quantitative eligibility criteria in T2DM studies: age, HbA1, and BMI, which are used in the free-text eligibility criteria of 49.1%, 48.5%, and 43.98% T2DM trials, respectively [13]. Age is also a required field in the eligibility criteria section of a trial summary in ClinicalTrials.gov. We therefore used the structured age field. To profile the target population of T2DM, we identified all the T2DM patients in the OneFlorida Data Trust using both ICD-9-CM and ICD-10-CM codes. To validate the findings of US-based T2DM trials on UK-based trials, we identified T2DM patients from the CALIBER data with the ICD-10 codes in the hospital data and the Readcodes in the primary care data. We applied the GIST metric to assess the *a priori* generalizability of T2DM trials in the US and UK using the target populations derived from OneFlorida Data Trust and CALIBER, respectively. We also stratified the analysis by trial phase. We assessed the *a posteriori* generalizability by comparing the characteristics of the enrolled patients reported in ClinicalTrials.gov with the T2DM patients in OneFlorida Data Trust and CALIBER, respectively.

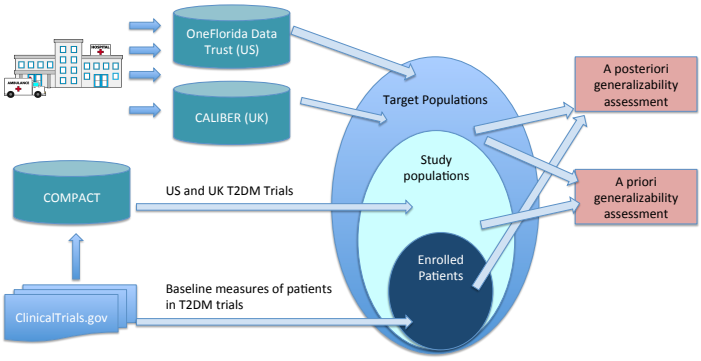


**Figure 1**. Analytical workflow of this study

***Dataset Preparation***

**1) Processing clinical trial summaries**. From the COMPACT database, we identified interventional clinical trials on T2DM with a study start date between January 2005 and September 2016. There are 1,671 such studies in the United States and 209 such studies in the United Kingdom.

**2) Identifying T2DM Patients in OneFlorida Data Trust**. Following existing literature [12], we identified patients with T2DM in OneFlorida Data Trust using the following criteria, where the patient (1) needs to have at least two diagnoses of Type 2 diabetes; (2) needs NOT have any Type 1 diabetes diagnoses; and (3) should have at least one HbA1c measurement regardless of their temporal relationships to diagnosis times. Diagnoses of Type 1 and 2 diabetes were identified with ICD-9-CM and ICD-10-CM diagnosis codes.

**3) Identifying T2DM Patients in CALIBER**. We utilized descriptive data from a previous study [31] using a deterministic to identify patients with Type 2 diabetes in CALIBER using diagnostic codes (Read codes in primary care, ICD-10 in secondary care).

***Assessing a Priori Generalizability***

To quantify the population representativeness of studies based on a single quantitative criterion (i.e., age, HbA1c, and BMI in this study), we calculated the univariate GIST scores for trial sets of different study phases [12]. The GIST score is the sum across all consecutive non-overlapping value intervals of the percentage of studies that recruit patients in that interval, multiplied by the percentage of patients observed in that interval:

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

where $N$ is the number of distinct value intervals of the quantitative feature, $T$ is the number of trials, $P$ is the number of patients, $w_j$ 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 low and high boundary values, 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 falls within the $i^{th}$ interval. The GIST score ranges from 0 to 1, with 0 being not generalizable and 1 being perfectly generalizable. It characterizes the proportion of patients potentially eligible across trials. For more detailed explanation of GIST, see [12, 14]. We have previously evaluated the validity of the GIST metric in quantifying the population representativeness of trials using simulated patient populations [32]. Compared with mGIST which gives an overall score for multiple variables, GIST gives variable-specific scores, allowing us to compare them with the results of *a posteriori* generalizability assessment which are also variable-specific.

***Assessing a Posteriori Generalizability***

To compare enrolled patients with the target population, we identified all the US and UK T2DM trials that have reported results in ClinicalTrials.gov and extracted the baseline measures including the number of participants, their gender and race, and the mean and standard deviation (SD) values of the three major measures: age, BMI, and HbA1c. We aggregated the mean and SD for age, HbA1C, and BMI separately for all the trials that report both mean and SD for the variable using the following formulas [33]:

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

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

where T is the number of studies. We assessed the *a posteriori* generalizability of the trials by comparing the aggregate mean and SD values of the three quantitative variables (i.e., age, HbA1c, and BMI) as well as the gender and race distributions with the real world population of T2DM patients in OneFlorida and CALIBER. We used the two-sample t-test to assess differences in quantitative variables (i.e., age, HbA1c, and BMI) and chi-square test to assess difference in categorical variables (i.e., race and gender) between the target population and enrolled patients.

**Results**

***Basic Characteristics of T2DM Trials***

Basic characteristics of the clinical trials on T2DM included in our analysis are shown in Table 1. Even though the number of trials differed significantly between the US and the UK, they exhibit similar characteristics. The rates of missing data for study phase are 23.6% and 27.3% in trials in the US and the UK respectively. A majority of trials are sponsored by industry (63.4% and 64.1%). Drugs were the most common interventions (72.1% and 72.7%),

followed by behavioral interventions (11.6% and 8.1%). Treatment was the primary purpose for the majority of the trials (75.9% and 77.0%). Most of the trials were randomized (86.9% and 90.1%).

**Table 1.** Characteristics of T2DM trials in the US and the UK

| **Study Characteristics** | **# of US-Based Trials (%) (N = 1,671)** | **# of UK-Based Trials (%) (N = 209)** |
|---|---|---|
| ***Study phase*** | -- | -- |
| Phase 0 | 9 (0.5%) | 0 (0%) |
| Phase 1 | 299 (17.9%) | 22 (10.5%) |
| Phase 2 | 349 (20.9%) | 23 (11.0%) |
| Phase 3 | 495 (29.6%) | 77 (36.8%) |
| Phase 4 | 189 (11.3%) | 33 (15.8%) |
| Unspecified | 394 (23.6%) | 57 (27.3%) |
| ***Sponsor type*** | -- | -- |
| NIH | 23 (1.4%) | 0 (0%) |
| Industry | 1,060 (63.4%) | 134 (64.1%) |
| Other U.S. Federal Agency | 30 (1.8%) | 0 (0%) |
| Other | 558 (33.4%) | 75 (35.9%) |
| ***Intervention type*** | -- | -- |
| Drug | 1,204 (72.1%) | 152 (72.7%) |
| Procedure | 34 (1.9%) | 2 (1.0%) |
| Biological | 32 (1.9%) | 4 (1.9%) |
| Device | 49 (2.9%) | 9 (4.3%) |
| Behavioral | 194 (11.6%) | 17 (8.1%) |
| Dietary supplement | 49 (2.9%) | 14 (6.7%) |
| Genetic | 1 (0.1%) | 0 (0%) |
| Radiation | 1 (0.1%) | 0 (0%) |
| Other | 107 (6.4%) | 11 (5.3%) |
| ***Primary purpose*** | -- | -- |
| Basic science | 108 (6.5%) | 14 (6.7%) |
| Diagnostic | 20 (1.2%) | 2 (2.0%) |
| Education/Counseling/Training | 1 (0.1%) | 0 (0%) |
| Health services research | 40 (2.4%) | 3 (1.4%) |
| Prevention | 99 (5.9%) | 17 (8.1%) |
| Screening | 4 (0.2%) | 0 (0%) |
| Supportive care | 33 (2.0%) | 5 (2.4%) |
| Treatment | 1,268 (75.9%) | 161 (77.0%) |
| Unspecified | 98 (5.9%) | 7 (3.3%) |
| ***Allocation*** | -- | -- |
| Randomized | 1,452 (86.9%) | 190 (90.1%) |
| Non-Randomized | 107 (6.4%) | 3 (1.4%) |
| Unspecified | 112 (6.7%) | 16 (7.7%) |

### *A Posteriori Generalizability of T2DM Trials*

Among all the T2DM trials in our analysis, 428 (25.6%, out of 1,671) US trials and 86 (41.1%, out of 209) UK trials report summary-level results (e.g., baseline characteristics and outcome measures) in ClinicalTrials.gov. Table 2 illustrates the number of T2DM trials that report mean values and standard deviation values for age, HbA1c, BMI, as well as race and gender. Within the trials that provided these statistics, most of them provided the mean and standard deviation of age for all the enrolled patients. A higher percentage of UK-based trials reported mean and standard deviation values of HbA1c and BMI than the US-based trials. The primary reasons that the remaining trials did not report any results in ClinicalTrials.gov include (1) still under recruitment, (2) completed before December 6, 2007 and thus not required to submit results, and (3) pending results (results of applicable trials of FDA-regulated drugs, biologic, and device must be submitted within 12 months of trial completion [34]).

**Table 2.** Number of T2DM trials that reported results in ClinicalTrials.gov

| Results | # of Trials in the US / Total # (%) (N = 1,671) | # of Trials in the UK / Total # (%) (N = 209) |
|---|---|---|
| # of trials with any results | 428/1,671 (25.6%) | 86/209 (41.1%) |
| # of trials reporting mean value of age | 400/428 (93.5%) | 80/86 (93.0%) |
| # of trials reporting standard deviation of age | 388/428 (90.7%) | 79/86 (91.9%) |
| # of trials reporting mean value of HbA1c | 131/428 (30.6%) | 38/86 (44.2%) |
| # of trials reporting standard deviation of HbA1c | 128/428 (29.9%) | 38/86 (44.2%) |
| # of trials reporting mean value of BMI | 109/428 (25.5%) | 30/86 (34.9%) |
| # of trials reporting standard deviation of BMI | 104/428 (24.3%) | 30/86 (34.9%) |
| # of trials reporting race | 159 /428(37.1%) | 34/86 (39.5%) |
| # of trials reporting gender | 426/428 (99.5%) | 86/86 (100.0%) |

We extracted the baseline characteristics of the patients enrolled in T2DM trials in the trial summaries on ClinicalTrials.gov. Table 3 reports the number of trials that provide the results of each baseline characteristic. We used formula (2) and (3) to calculate the weighted mean and standard deviation values for quantitative variables (i.e., age, BMI, and HbA1c) separately.

**Table 3.** Number of T2DM Trials that provided the results for the baseline measures

| Reported Results | # of US-Based T2DM Trials (# of patients) | # of UK-Based T2DM Trials (# of patients) | Two-Tailed *P* Values |
|---|---|---|---|
| Any baseline measures | 428 (193,345) | 86 (90,026) | -- |
| Mean and standard deviation of age | 388 (185,560) | 79 (87,441) | $P < 0.0001$ |
| Mean and standard deviation of HbA1c | 128 (65,445) | 38 (32,314) | $P < 0.0001$ |
| Mean and standard deviate n of BMI | 104 (68,515) | 30 (36,768) | $P < 0.0001$ |
| Race (white, black, Asian, other) | 159 (96,518) | 34 (53,016) | $P < 0.0001$ |
| Gender (male, female) | 426 (192,721) | 86 (89,791) | $P < 0.0001$ |

Table 4 compares the characteristics of the patients enrolled in T2DM trials and the T2DM patients in the OneFlorida Data Trust and CALIBER. The differences of mean values of age between the patients in OneFlorida and the patients enrolled in US-based T2DM trials of Phase I, II, II and are 21.9, 13.0, and 11.0, respectively, showing an increasing *a posteriori* generalizability of US-based trials regarding age. The patients enrolled in both US and UK-based T2DM trials were younger than the T2DM patients in the clinical data warehouses (two-tailed $p < 0.0001$). Regarding race, Caucasian/White and Asian were overrepresented, whereas women, black, and other races were underrepresented. The differences between the target populations and the enrolled patients in US and UK-based trials are both statistically significant with respect to race (two-tailed $p < 0.0001$). Regarding gender, female patients were underrepresented in both US-and UK-based trials (two-tailed $p < 0.0001$). The patients enrolled in both US and UK-based trials have higher HbA1c values than the target populations (two-tailed $p < 0.0001$). The patients enrolled in US-based trials have a slightly lower BMI (two-tailed $p < 0.0001$), whereas the patients enrolled in UK-based trials have a slightly higher BMI (two-tailed $p < 0.0001$).

**Table 4.** Characteristics of patients enrolled in T2DM trials and T2DM patients in OneFlorida Data Trust and CALIBER

| Characteristic | United States | | | United Kingdom | | |
|---|---|---|---|---|---|---|
| | T2DM Patients in OneFlorida Data Trust | Patients in US-Based T2DM Trials | Difference[a] | T2DM Patients in CALIBER | Patients in UK-Based T2DM Trials | Difference |
| Total # of T2DM patients | 148,970 | -- | -- | 150,665 | -- | -- |
| Age (mean ± SD) | 68.5 ± 14.2 | 58.1 ± 9.6 | -10.4 | 64.9 ± 13.86 | 60.2 ± 9.4 | -4.7 |
| *Race* | 148,970 | 96,518 | -- | 57,192 | 53,016 | -- |
| Caucasian / White (%) | 88,213 (59.2%) | 72,647 (75.3%) | 16.1% | 28,629 (50.1%) | 39,806 (75.1%) | 25.0% |
| Black (%) | 40,186 (27.0%) | 6,604 (6.8%) | -20.2% | 2,757 (9.6%) | 2,979 (5.6%) | -4.0% |
| Asian (%) | 2,594 (1.7%) | 10,323 (10.7%) | 9.0% | 5,018 (8.8%) | 6,511 (12.3%) | 3.5% |
| Other (%) | 17,977 (12.1%) | 6,825 (7.1%) | -5.0% | 20,788 (36.3%) | 3,710 (7.0%) | -29.3% |

| | | | | | | |
|---|---|---|---|---|---|---|
| *Gender* | 148,969 | 192,721 | -- | 150,658 | 89,791 | -- |
| Male (%) | 69,221 (46.5%) | 106,821 (55.4%) | 8.9% | 81,312 (54.0%) | 53,703 (59.8%) | 5.8% |
| Female (%) | 79,748 (53.5%) | 85,900 (44.6%) | -8.9% | 69,346 (46.0%) | 36,088 (40.2%) | -5.8% |
| BMI, $kg^2/m$ (mean ± SD) | 32.3 ± 8.2 | 31.7 ± 5.5 | -0.6 | 29.3 ± 6.0 | 30.9 ± 5.6 | 1.6 |
| HbA1c, % (mean ± SD) | 7.5 ± 2.9 | 8.2 ± 1.0 | 0.7 | 7.8 ± 1.8 | 8.0 ± 1.0 | 0.2 |

[a] The difference between the patients enrolled in the trials and the patients in the target population.

***Population Representativeness of T2DM Trials***

Figure 2 visualizes the target population, study population, and the enrolled patients in the US and UK-based T2DM trials for each of the three major quantitative eligibility criteria: age, BMI, and HbA1c. The green curves represent the study populations of the eligible patients, i.e., the percentage of trials that allow a certain value of the variable. The red curves represent the distribution of patients enrolled in the T2DM trials over the value spectrum of the variable. The blue curves represent the target population. In general, the visualization of these populations for US and UK trials exhibited similar characteristics. For example, there is a noticeable gap between the patients enrolled in the T2DM trials and the target population with respect to all the three quantitative criteria. The patients enrolled in T2DM trials were younger and had lower HbA1c and BMI values than those in the target populations (red curves vs. blue curves). The T2DM trials usually permit patients of a wide range of age values. The gap between the study population and the target population is obvious for HbA1c (green curves vs. blue curves). The trends of the study populations and enrolled patients are similar (green curves vs. red curves).

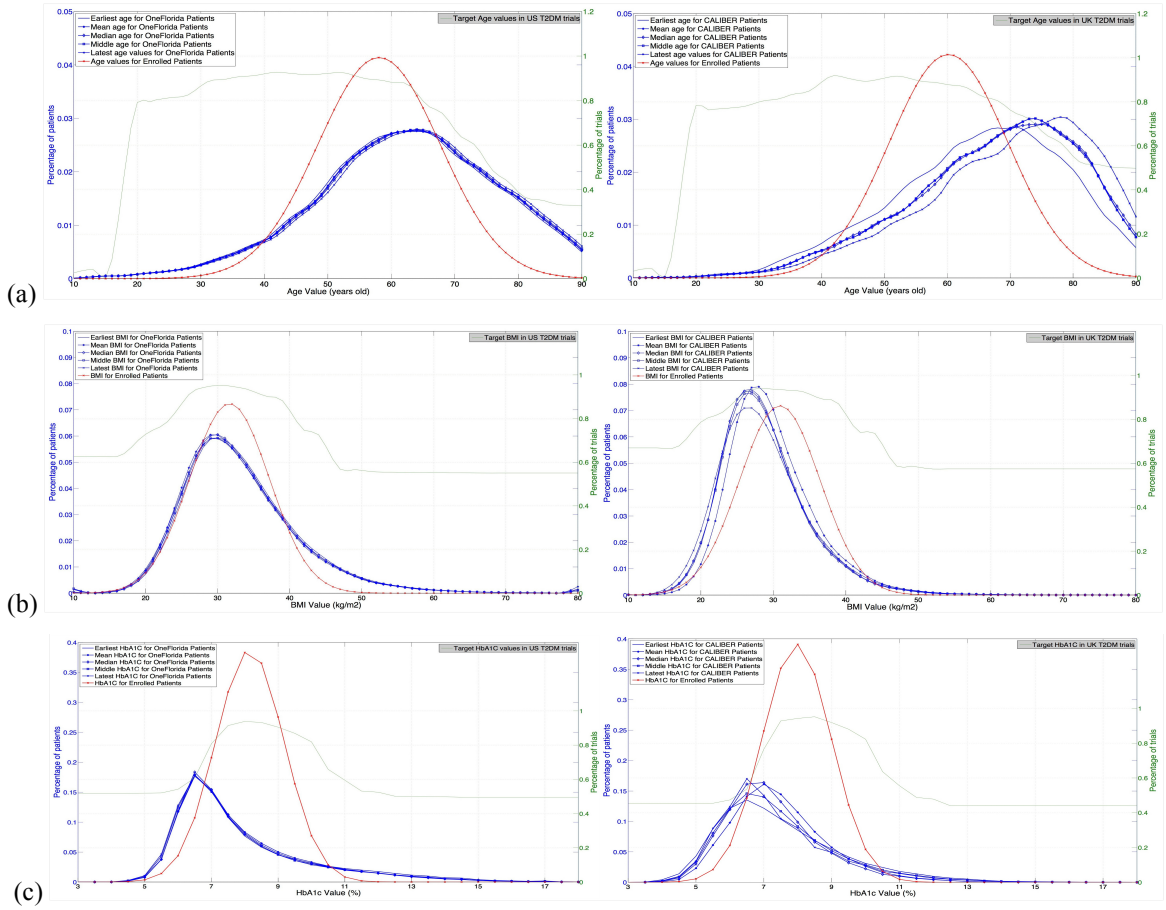


**Figure 2.** Visualization of (a) age, (b) BMI, and (c) HbA1c in the target populations and study populations of US and UK-based T2DM trials, respectively.

We used the GIST metric to assess the population representativeness of US and UK-based T2DM trials with respect to the three major quantitative eligibility criteria: age, BMI, and HbA1c (Table 5). The GIST scores for the US and UK-based trials were calculated using the target populations derived from the OneFlorida Data Trust data and CALIBER, respectively. We stratified the analysis by study phase. As shown in Table 5, US and UK trials have

similar overall GIST scores for all three variables. The GIST scores of age in both the US and UK trials increased from Phase I to Phase III, which is consistent with the results from two previous studies using the T2DM patients data in the Columbia University [12] and a national survey [14] as the target populations. It is also consistent with our finding of the *a posteriori* generalizability assessment conducted in this work. Phase I studies in the UK had lower GIST scores of age than Phase I studies in the US. The GIST scores of 0.27-0.36 indicate a serious population representativeness issue of UK T2DM Phase I trials. However, Phase II trials in the UK had higher GIST scores of age than Phase II trials in the US. With respect to BMI, the UK trials had slightly higher GIST scores than the US trials of all phases. With respect to HbA1c, the GIST scores of the US trials decreased from Phase I to Phase III, which is also consistent with the two previously mentioned studies [12, 14]. The HbA1c's GIST scores of the UK trials decreased from Phase I to Phase II, while the UK trials had similar GIST scores of HbA1c as US trials. The fact that *a priori* generalizability of HbA1c is the lowest among the three criteria is consistent with the visualization shown in Figure 2, as well as the *a posteriori* generalizability assessment.

**Table 5.** GIST scores of age, BMI, and HbA1c of T2DM trials in different phases.

| **Variable** | **Reading** | **US-Based T2DM Trials** | | | | **UK-Based T2DM Trials** | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | All | Phase I | Phase II | Phase III | All | Phase I | Phase II | Phase III |
| | N | 1,671 | 299 | 349 | 495 | 209 | 22 | 23 | 77 |
| Age | Earliest | 0.74 | 0.55 | 0.76 | 0.87 | 0.76 | 0.36 | 0.84 | 0.89 |
| | Mean | 0.74 | 0.55 | 0.75 | 0.86 | 0.74 | 0.31 | 0.80 | 0.88 |
| | Median | 0.74 | 0.54 | 0.75 | 0.86 | 0.74 | 0.31 | 0.80 | 0.88 |
| | Middle | 0.73 | 0.54 | 0.75 | 0.86 | 0.74 | 0.31 | 0.81 | 0.88 |
| | Latest | 0.73 | 0.53 | 0.74 | 0.86 | 0.71 | 0.27 | 0.77 | 0.87 |
| BMI | Earliest | 0.87 | 0.78 | 0.87 | 0.88 | 0.89 | 0.79 | 0.91 | 0.94 |
| | Mean | 0.87 | 0.79 | 0.87 | 0.89 | 0.89 | 0.76 | 0.93 | 0.94 |
| | Median | 0.87 | 0.79 | 0.87 | 0.89 | 0.89 | 0.79 | 0.91 | 0.94 |
| | Middle | 0.87 | 0.79 | 0.87 | 0.88 | 0.89 | 0.78 | 0.91 | 0.93 |
| | Latest | 0.87 | 0.79 | 0.86 | 0.88 | 0.88 | 0.78 | 0.90 | 0.93 |
| HbA1c | Earliest | 0.73 | 0.83 | 0.71 | 0.64 | 0.69 | 0.81 | 0.61 | 0.62 |
| | Mean | 0.74 | 0.84 | 0.72 | 0.65 | 0.74 | 0.84 | 0.68 | 0.69 |
| | Median | 0.73 | 0.84 | 0.72 | 0.65 | 0.71 | 0.83 | 0.65 | 0.64 |
| | Middle | 0.73 | 0.84 | 0.71 | 0.64 | 0.70 | 0.82 | 0.63 | 0.63 |
| | Latest | 0.73 | 0.84 | 0.71 | 0.64 | 0.69 | 0.82 | 0.61 | 0.61 |

**Discussion and Conclusions**

In this study, we used real-world patient data in the target population to assess the generalizability of T2DM clinical trials in the US and UK. As shown in Table 5, US and UK-based T2DM trials have similar *a priori* generalizability of age, HbA1c, and BMI. However, the GIST scores for age in trials of different phases differ between US and UK-based trials. While GIST provides a quantitative metric for comparing the population representativeness of different sets of trials, visualization of different populations can reveal the systematically omitted or overly included population subgroups. The results of *a priori* generalizability showed that the US and UK-based trials exhibit similar issues with respect to the three most frequently used quantitative criteria. The results of the *a posteriori* generalizability showed that males, whites, and Asians are overrepresented in both US and UK-based T2DM trials while females, blacks, and other races are underrepresented.

Compared to the *a posteriori* generalizability assessment, the use of the GIST metric to assess the *a priori* generalizability assessment has a few advantages. First, it can be performed during the trial design phase, which would help reveal issues of eligibility criteria that are biased towards certain population subgroups, and help trial designers optimize the balance between internal and external validity. Trial designers can fine-tune the criteria without diminishing the internal validity. For example, UK T2DM Phase I trials have a very low GIST score of age. Trial designers should thus adjust the restrictive age criterion in Phase I trials in the future to improve their population representativeness. Second, GIST quantifies the difference of the distributions of the eligible patients and the target population over a variable, whereas the *a posteriori* generalizability compares the mean difference of a variable. It can be done using cost-effective informatics tools. Meanwhile, the *a priori* generalizability assessment has a few disadvantages. First, it does not take into account the practical issues in the trial recruitment phase, such as

geographic locations, accessibility of trial information, and consideration of comorbidities. For example, due to real-world complications, most studies failed to recruit representative samples of their study population as specified in trial eligibility criteria [10]. On the other hand, the *a posteriori* generalizability assessment, which compares the enrolled patients with the target population, can provide a more accurate assessment of the population representativeness. The issues of gender disparity and race disparity, as well as other disease-specific outcome measures can be accurately detected. Both *a priori* and *a posteriori* generalizability results showed increasing generalizability of US-based T2DM trials from Phase I to Phase II with respect to age, which confirmed our hypothesis that *a priori* generalizability is comparable to *a posteriori* generalizability in identifying certain restrictive quantitative eligibility criteria. It should be a common practice to assess both *a priori* generalizability based on trial design factors such as eligibility criteria before patient recruitment as well as *a posteriori* generalizability *post hoc* based on enrolled patients. The trial design issues that are found in *a priori* generalizability can be addressed before patient recruitment, thereby improving the *a posteriori* generalizability and the cost-benefit ratio of the trials. Nevertheless, clinical trial investigators should also consider practical issues in the trial recruitment phase. For example, most trials still use a traditional hospital-based recruitment strategy. Thus, the trial designers of these studies should carefully choose recruitment sites, and take into account the population characteristics of these sites' catchment areas.

A number of limitations should be noted in our study. First, less than 40% of T2DM trials reported results in ClinicalTrials.gov. Therefore, the aggregate results of the enrolled patients represent merely a convenient sample. Second, the OneFlorida Data Trust contains data of patients who have visited healthcare organizations in the state of Florida, and thus might not be representative. Even though we only used patient data from the state of Florida in the US, Florida is the third most populous state (~19.9 million) in the US. Third, it is possible that not all clinical trials conducted in the UK are registered on ClinicalTrials.gov. We identified trials based on the study site. Some trials are conducted in multiple countries. Therefore, some UK-based trials may also have study sites in other countries.

**Acknowledgments**

We show our gratitude to Sarah Mixon at Florida State University Department of Chemistry for her hard work to extract results from Type 2 diabetes trials in the United Kingdom. The development of the COMPACT database and the VITTA tool was supported by U.S. National Library of Medicine Grant R01LM009886 (PI: Weng) and the National Center for Advancing Translational Science (NCATS) Award UL1TR000040 (PI: Ginsberg). This work was partially supported by an Amazon Web Service in Education Research Grant Award (PI: He) and the Planning Grant of the Institute of Successful Longevity at Florida State University. The work was also partially supported by National Center for Advancing Translational Sciences under the Clinical and Translational Science Award UL1TR001427 (PI: Nelson & Shenkman). The content is solely the responsibility of the authors and does not represent the official view of the National Institutes of Health.

## References

1. Fahey T, Griffiths S, Peters TJ. Evidence based purchasing: understanding results of clinical trials and systematic reviews. BMJ. 1995;311(7012):1056-9; discussion 9-60.
2. Leaf C. Do Clinical Trials Work? The New York Times. 2013.
3. Sardar MR, Badri M, Prince CT, Seltzer J, Kowey PR. Underrepresentation of women, elderly patients, and racial minorities in the randomized trials used for cardiovascular guidelines. JAMA Intern Med. 2014;174(11):1868-70. PMID: 25264856.
4. Lewis JH, Kilgore ML, Goldman DP, Trimble EL, Kaplan R, Montello MJ, et al. Participation of patients 65 years of age or older in cancer clinical trials. J Clin Oncol. 2003;21(7):1383-9.
5. 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. PMID: 15380160.
6. Cigolle CT, Blaum CS, Halter JB. Diabetes and cardiovascular disease prevention in older adults. Clin Geriatr Med. 2009;25(4):607-41, vii-viii. PMID: 19944264.
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):dju05. PMID: 24647464.
8. Cahan A, Cahan S, Cimino JJ. Computer-aided assessment of the generalizability of clinical trial results. Int J Med Inform. 2017;99:60-6.
9. Reichert JM. Trends in development and approval times for new therapeutics in the United States. Nat Rev Drug Discov. 2003;2(9):695-702. PMID: 12951576.

10. 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. PMID: 18557666.
11. Weng C. Optimizing Clinical Research Participant Selection with Informatics. Trends Pharmacol Sci. 2015;36(11):706-9.
12. 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.
13. He Z, Carini S, Sim I, Weng C. Visual aggregate analysis of eligibility features of clinical trials. J Biomed Inform. 2015;54:241-55.
14. He Z, Wang S, Bornanian E, Weng C. Assessing the Population Representativeness of Type 2 Diabetes Trials by Combining Public Data from ClinicalTrials.gov and NHANES. Stud Health Technol Inform. 2015;2015(216):569-73. PMID: 26262115.
15. He Z, Ryan P, Hoxha J, Wang S, Carini S, Sim I, et al. Multivariate analysis of the population representativeness of related clinical studies. J Biomed Inform. 2016;60:66-76. PMID: 26820188.
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. WHO. Global Report on Diabetes. World Health Organization [February 12, 2017]. Available from: http://apps.who.int/iris/bitstream/10665/204871/1/9789241565257_eng.pdf?ua=1.
18. Association AD. Diabetes complications [February 12, 2017]. Available from: http://www.diabetes.org/living-with-diabetes/complications/.
19. WebMD. Diabetes: Differences Between Type 1 and 2 - Topic Overview. Available from: http://www.webmd.com/diabetes/tc/diabetes-differences-between-type-1-and-2-topic-overview.
20. ADA. We Are Research Leaders [February 12, 2017]. Available from: http://www.diabetes.org/research-and-practice/we-are-research-leaders.
21. Diabetes UK: Researh Round-Up [February 12, 2017]. Available from: https://www.diabetes.org.uk/Research/Research-round-up/.
22. OneFlorida. OneFlorida: Clinical Research Consortium [March 18, 2016]. Available from: http://onefloridaconsortium.org.
23. Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol. 2012;41(6):1625-38.
24. Zarin DA, Tse T, Williams RJ, Carr S. Trial Reporting in ClinicalTrials.gov - The Final Rule. N Engl J Med. 2016;375(20):1998-2004.
25. Ogino D, Takahashi K, Sato H. Characteristics of clinical trial websites: information distribution between ClinicalTrials.gov and 13 primary registries in the WHO registry network. Trials. 2014;15:428.
26. Hao T, Weng C. Valx - Numerical Expression Extraction and Normalization [August 1, 2014]. Available from: http://columbiaelixr.appspot.com/valx.
27. Hao T, Liu H, Weng C. Valx: A System for Extracting and Structuring Numeric Comparison Statements from Text. Methods of Information in Medicine. 2016;In press.
28. Miotto R, Weng C. Unsupervised mining of frequent tags for clinical eligibility text indexing. J Biomed Inform. 2013;46(6):1145-51.
29. 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;2014:1777-86.
30. Morley KI, Wallace J, Denaxas SC, Hunter RJ, Patel RS, Perel P, et al. Defining disease phenotypes using national linked electronic health records: a case study of atrial fibrillation. PLoS One. 2014;9(11):e110900.
31. Dinesh Shah A, Langenberg C, Rapsomaniki E, Denaxas S, Pujades-Rodriguez M, Gale CP, et al. Type 2 diabetes and incidence of a wide range of cardiovascular diseases: a cohort study in 1.9 million people. Lancet. 2015;385 Suppl 1:S86.
32. He Z, Chandar P, Ryan P, Weng C. Simulation-based Evaluation of the Generalizability Index for Study Traits. AMIA Annu Symp Proc. 2015;2015:594-603.
33. Wikipedia. Pooled Variance [cited October 19, 2014]. Available from: http://en.wikipedia.org/wiki/Pooled_variance.
34. FDAAA 801 Requirements [February 16, 2017]. Available from: https://clinicaltrials.gov/ct2/manage-recs/fdaaa.