Journal of Biomedical Informatics 63 (2016) 325–336

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

GIST 2.0: A scalable multi-trait metric for quantifying population representativeness of individual clinical studies

Anando Sen [a], Shreya Chakrabarti [a], Andrew Goldstein [a,b], Shuang Wang [c], Patrick B. Ryan [a,d], Chunhua Weng [a,*]

[a] Department of Biomedical Informatics, Columbia University, New York, NY 10032, United States
[b] Department of Medicine, New York University, New York NY 10016, United States
[c] Department of Biostatistics, Columbia University, New York, NY 10032, United States
[d] Janssen Research and Development, Titusville, NJ 08560, United States

ARTICLE INFO

Article history:
Received 24 March 2016
Revised 2 July 2016
Accepted 2 September 2016
Available online 4 September 2016

Keywords:
Clinical trials
Generalizability
Population representativeness
Trait dependencies
Eligibility criteria

ABSTRACT

The design of randomized controlled clinical studies can greatly benefit from iterative assessments of population representativeness of eligibility criteria. We propose a multi-trait metric - GIST 2.0 that can compute the *a priori* generalizability based on the population representativeness of a clinical study by explicitly modeling the dependencies among all eligibility criteria. We evaluate this metric on twenty clinical studies of two diseases and analyze how a study's eligibility criteria affect its generalizability (collectively and individually). We statistically analyze the effects of trial setting, trait selection and trait summarizing technique on GIST 2.0. Finally we provide theoretical as well as empirical validations for the expected properties of GIST 2.0.

1. Introduction

Randomized controlled trials generate medical evidence of the highest quality. Hence it is of great importance that clinical studies benefit a representative proportion of the population under consideration. The representativeness of a trial affects its generalizability [1–4], which indicates whether the findings of a trial can be extended to patients of the same disease who are not trial participants but for whom the treatment is intended. This becomes of prime importance when the results of a clinical trial are disseminated to other patients since the lack of generalizability can lead to serious negative consequences in some subgroups of the diseased population that may have been underrepresented in the trial. A key contributing factor for generalizability is the trial eligibility criteria, which define constraints on the various study traits. Study traits are attributes of a patient that are relevant for the study (either to determine eligibility or to measure outcome). They include conditions (e.g. type 2 diabetes), procedures (e.g. colonoscopy), medications (e.g. metformin), laboratory tests (e.g. glucose) or demographic information (e.g. ethnicity).

Inappropriate eligibility criteria can result in studies that either exclude patients who might benefit from the intervention or, conversely, threaten patient safety by causing unforeseeable post-marketing adverse drug effects [5–7]. Resources that assist clinical investigators make better eligibility criteria choices are very limited. Clinical research eligibility criteria often suffer from ambiguity, complexity or over-restrictiveness [8,9]. The lack of interoperability with different data sources is another concern with current eligibility criteria [4]. Study designers often reuse eligibility criteria from previous clinical trials with minimal modifications [10], which may lead to a concordant bias in sampling and under-representation of certain subgroups. Some other researchers rely on past experience for patient selection. However, this type of selection process is highly subjective with limited justification [11]. Another popular practice in eligibility criteria design is through trial and error, which can be unstable and can entail frequent and costly protocol amendments. Hence, optimization of eligibility criteria is a topic of great interest.

Abbreviations: GIST, generalizability index for study traits; EHR, electronic health record; TP, target population; EP, EHR population; SP, study population; SS, study sample; sGIST, single-trait GIST; mGIST, multiple-trait GIST; T2DM, type 2 diabetes mellitus; IDA, iron deficiency anemia; HbA1C, glycated hemoglobin; ANOVA, analysis of variance.

* Corresponding author.
E-mail addresses: as5050@cumc.columbia.edu (A. Sen), sc4025@cumc.columbia.edu (S. Chakrabarti), ag3304@cumc.columbia.edu (A. Goldstein), sw2206@cumc.columbia.edu (S. Wang), ryan@ohdsi.org (P.B. Ryan), cw2384@cumc.columbia.edu (C. Weng).

http://dx.doi.org/10.1016/j.jbi.2016.09.003
1532-0464/

To address these concerns with eligibility criteria, we propose a metric to calculate *a priori* generalizability of a single trial based on its population representativeness. Currently, the lack of population representativeness in clinical studies remains largely undiscovered until after study publications (e.g. [12] - details in Section 1.2). With our metric we aim to provide a decision aid for eligibility criteria designs by answering important questions such as: (1) Are the eligibility criteria too restrictive (when multiple traits are considered together)? (2) Is there a particular eligibility trait (or traits) that decreases the study's population representativeness? (3) How would small changes in the eligibility criteria affect overall population representativeness of the study? The answers to these questions can potentially optimize population representativeness of the study within the constraints of patient safety and other Food and Drug Administration (FDA) regulations [13].

1.1. Populations in a clinical trial

For any clinical trial, there are typically four associated populations. The target population (TP) corresponds to the entire universe of patients (suffering from the disease under consideration) for whom the results of the clinical trial are intended. It includes patients who are unaware of the presence of the disease and those who do not seek medical treatment. The Electronic Health Record population (EP) includes those patients (suffering from that disease) who visit medical facilities to receive treatment and consultation from clinicians. A study population (SP) is the set of all patients who satisfy the eligibility criteria of a particular trial. Finally, the patients who actually enroll for the clinical trial constitute the study sample (SS).

The relationships among these populations are shown schematically in Fig. 1. The TP subsumes all other populations. It is impossible to characterize this population exactly, but it can often be approximated by the EP [14]. The dashed outline around the TP in Fig. 1 indicates that the TP is not exactly defined. The SP is determined solely by eligibility criteria, and may exclude specific subpopulations (e.g. elderly patients, children, patients with comorbidities etc.) who may potentially benefit from the trial. The SP subsumes the SS but both these populations may include patients from outside the EP. In the scenario of perfect recruitment (the SS being a random sample from the SP) all subgroups within the SP are represented in the SS. However, this may not be the case as the SS is constrained by informed consent, geographical locations, ability to adhere to the conditions set by the trial, etc.

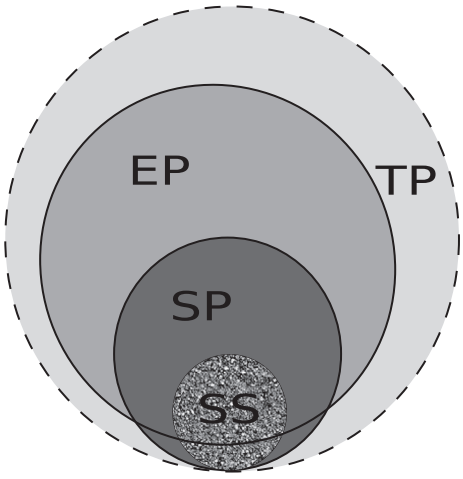

Fig. 1. Relationship between various populations associated with a clinical trial: target population (TP), EHR population (EP), study population (SP) and study sample (SS).

Generalizability can be measured before the commencement of a trial (*a priori*) or after its completion (*a posteriori*). *A priori* generalizability is calculated on the basis of eligibility criteria and we refer to this as eligibility-driven generalizability. Since the SP is defined precisely by the eligibility criteria, this is measuring the representativeness of the SP within the TP. *A posteriori* generalizability is determined by the actual patients enrolled in the study, i.e. the SS. This is sample-driven generalizability, which measures the representativeness of the SS within the TP. For this paper we focus on the former, which is affected by population representativeness.

1.2. Previous work and its limitations

A detailed literature review for clinical trial generalizability was discussed by Kennedy-Martin et al. [15]. As mentioned above, most of the generalizability assessments have been performed after the completion of a trial. For example, in a technical report by Buchanan et al. [12], a generalizability study was performed for HIV treatment clinical trials. The majority of the results presented in this study were simulation-based and only two clinical trials were evaluated. Bress et al. studied the generalizability of the Systolic Blood Pressure Intervention Trial (SPRINT) in detail [16]. Although the analysis was comprehensive, it was limited to a single trial. The concept of *a priori* generalizability has been mentioned by several authors but there have been relatively few efforts at a rigorous quantitative assessment. Such assessments have been restricted to visualization techniques (e.g. comparison of histograms by Schoenmaker et al. [5]) and statistical tests (e.g. assessment of generalizability bias [17,18]).

One of the first efforts at quantification of generalizability used receiver operator characteristic analysis [19]. A binary classifier evaluated infants with fever for presence of bacterial infection. Training and validation sets consisted of patients from two time periods in different hospitals. This type of supervised study is different from a clinical trial as prior knowledge about the outcome (bacterial infection in this case) was known. Intervention outcomes cannot be assumed in clinical trial settings. Some studies have analyzed the generalizability of one trial in detail but their methods remain untested with a broader class of clinical studies. For example, (1) Wang et al. computed the representativeness of the 'Relaxin for the Treatment of Acute Heart Failure (RELAX-AHF) trial' by calculating the fraction of patients in international registries who would satisfy the eligibility criteria [20]. (2) Cole et al. [21] used inverse-probability selection weights to calculate the generalizability of the AIDS clinical trial group (ACTG) 320 trial (for HIV) to a target population (of all HIV patients in the USA) defined by state registries.

In most of the above cases, studies with multiple eligibility criteria had each criterion evaluated independently. However, there could be functional relationships between two or more traits. We refer to such relationships between traits as trait dependencies. Furthermore, every trait was treated as equally important in the computation of generalizability. In actual practice the importance of a trait may be disease-specific (e.g. HbA1C is more important in type two diabetes than it is in chronic kidney disease) as well as trial-specific (e.g. HbA1C >6.5% is less important than HbA1C within 9–11% - see Section 2.3). We refer to a quantification for the importance of a study trait as trait-significance (this should not be confused with 'statistical significance' later in the paper). Hence, the two major limitations of all the studies mentioned above were: (a) the trait dependencies were not explicitly modeled (b) the significance of traits was not accounted for. In addition, most of the studies (with the exception of [16]) restricted their generalizability analyses to continuous traits and did not consider categorical traits, which are often critical in clinical study designs.