Citation: Clin Transl Sci (2020) 13, 675–684; doi:10.1111/cts.12764

REVIEW

Clinical Trial Generalizability Assessment in the Big Data Era: A Review

Zhe He[1,*], Xiang Tang[2], Xi Yang[3], Yi Guo[3], Thomas J. George[4], Neil Charness[5], Kelsa Bartley Quan Hem[6], William Hogan[3] and Jiang Bian[3]

Clinical studies, especially randomized, controlled trials, are essential for generating evidence for clinical practice. However, generalizability is a long-standing concern when applying trial results to real-world patients. Generalizability assessment is thus important, nevertheless, not consistently practiced. We performed a systematic review to understand the practice of generalizability assessment. We identified 187 relevant articles and systematically organized these studies in a taxonomy with three dimensions: (i) data availability (i.e., before or after trial (*a priori* vs. *a posteriori* generalizability)); (ii) result outputs (i.e., score vs. nonscore); and (iii) populations of interest. We further reported disease areas, underrepresented subgroups, and types of data used to profile target populations. We observed an increasing trend of generalizability assessments, but < 30% of studies reported positive generalizability results. As *a priori* generalizability can be assessed using only study design information (primarily eligibility criteria), it gives investigators a golden opportunity to adjust the study design before the trial starts. Nevertheless, < 40% of the studies in our review assessed *a priori* generalizability. With the wide adoption of electronic health records systems, rich real-world patient databases are increasingly available for generalizability assessment; however, informatics tools are lacking to support the adoption of generalizability assessment practice.

Appropriately designed clinical research studies, especially randomized, controlled trials (RCTs), provide "gold-standard" evidence for determining the efficacy and safety of medical interventions,[1] allowing regulatory agencies to approve new therapies and care providers to make better clinical decisions. Nevertheless, trial investigators and sponsors often overemphasize the internal validity (i.e., "the extent to which observed treatment effects can be ascribed to differences in treatment and not confounding, thereby allowing the inference of causality to be ascribed to a treatment"[2]) of a study—rightfully to protect participants from undue harm and to collect sufficient efficacy information.[3] Typically excluded are pregnant women due to concern for fetal health[4] and patients with concomitant diseases to avoid noise in safety data.[5] However, overemphasis on internal validity can lead to exclusion of certain population subgroups and, subsequently, poor generalizability.[6] Unjustified exclusion of diverse and complex participants in clinical trials may undermine safety for patients who will use the drug in real-world settings.[5] Because of generalizability issues many approved drugs had been withdrawn from the market after severe adverse drug reactions (e.g., high toxicity, organ damage, and fatalities) were observed.[7]

The notions of generalizability and population representativeness are distinct but closely related. In clinical trials, three essential populations of interest exist: (i) the target population (TP)—patients to whom the study results are intended to be applied in real-world patients; (ii) the study population (SP)—patients who are eligible for the study (based on study inclusion/exclusion criteria); and (iii) the study sample (SS)—participants who are enrolled in the clinical study. Generalizability is the ultimate portability of the causal effects of an intervention (developed based on the SS) to the TP. Population representativeness—measuring the SP's coverage of the TP—is a key determining factor for generalizability. Other factors, such as variation of patients in different clinical settings, discrepancies in conditions under which a trial is conducted,[8] and incomplete reporting,[9] may also affect study generalizability. Further, many real-world constraints, such as trial awareness[10] and transportation,[11] can also affect participant enrollment. Thus, the SS may not adequately represent the SP and, subsequently, the TP.

In this review we focus on "population representativeness," and thus use the terms "population representativeness," "external validity," and "generalizability" interchangeably, omitting other extrinsic factors. A commonly used simplistic approach to assess generalizability is to assess the differences in patient characteristics between the study sample and the target population (i.e., patients who received the same treatment in routine care). Increasingly, approaches that compare the outcomes of patients from observational cohorts with participants in the original trials[12] were developed to evaluate study generalizability. However, these comparisons can only be made after trial completion.

Zhe He and Jiang Bian are co-corresponding authors.

[1]School of Information, Florida State University, Tallahassee, Florida, USA; [2]Department of Statistics, Florida State University, Tallahassee, Florida, USA; [3]Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, Florida, USA; [4]Hematology & Oncology, Department of Medicine, College of Medicine, University of Florida, Gainesville, Florida, USA; [5]Department of Psychology, Florida State University, Tallahassee, Florida, USA; [6]Calder Memorial Library, Miller School of Medicine, University of Miami, Miami, Florida, USA. *Correspondence: Zhe He (zhe@fsu.edu)
Received: November 20, 2019; accepted: January 25, 2020. doi:10.1111/cts.12764

More recently, another type of generalizability assessment method has emerged—making population comparisons based on data from study eligibility criteria and from observational cohorts generated through standard of care (e.g., electronic health records (EHRs)).[13] For example, one can compare eligible patients from an observational cohort (e.g., trial patients with stage IV colorectal cancer) with the target population of the study (e.g., all patients with stage IV colorectal cancer).

Generalizability assessment methods can be organized into two major categories based on whether the assessment data are available before or after trial completion: (i) the *a priori* (also called *eligibility-driven*) generalizability—the representativeness of *eligible* (study population) to the target population; and (ii) the *a posteriori* (or *sample-driven*) generalizability—the representativeness of *enrolled* participants (study sample) to the target population.

Although study generalizability is well-recognized, there is a significant knowledge gap between the methods and data available for generalizability assessment and their adoption in practice. To understand this gap, we performed a systematic review, identifying barriers and opportunities in clinical study generalizability assessment practice. To the best of our knowledge, only one previous review on generalizability was published—in 2015, before the emergence of quantitative, often informatics-based, *a priori* generalizability studies.[14] Further, our ultimate goal is to develop a decision tool to guide investigators on how to choose proper generalizability assessment methods for their clinical studies. Based on our review, we created a taxonomy that synthesizes existing generalizability assessment methods to inform the development of a decision guide. We also argue that, given the increasing availability of large-scale clinical data and advancements in informatics methods such as computable phenotypes, informaticians have an opportunity to develop novel generalizability assessment methods that could optimize patient selection in the study design phase.

IDENTIFICATION OF AVAILABLE INFORMATION

We performed a literature search over the following four databases: MEDLINE, Cochrane, PychINFO, and CINAHL. Following the Institute of Medicine's standards for systematic review[15] and Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA),[16] we conducted the review in six steps: (i) gaining an initial understanding about generalizability assessment and related concepts; (ii) identifying relevant keywords; (iii) formulating four search queries (see **Table S1** in **Supplementary File I**) to identify relevant articles; (iv) screening through titles and abstracts; (v) reviewing articles' full text to further filter out irrelevant articles; and (vi) coding the articles for data extraction.

Study selection and screening process

We used an iterative process to identify and refine the search keywords and strategies. Using the search strategies in **Table S1**, we identified 5,352 articles as of April 2019. After removing duplicates, 3,568 records were assessed for relevancy by two researchers (Z.H. and X.T.) through reviewing the titles and abstracts against the inclusion and exclusion criteria. Conflicts were resolved with a third reviewer (J.B.). During the screening process, we also iteratively refined the criteria (**Table 1**). Of the 3,568 articles, 3,275 were excluded through the abstract screening process. Subsequently, we reviewed the full texts of 293 articles, excluding 106 more articles based on the exclusion criteria. The interrater reliability of the full-text review is 0.90 (Cohen's kappa, $P < 0.001$).[17] One hundred eighty-seven articles were included in the final review. **Figure 1** is the PRISMA flow diagram that depicts the number of articles identified, included, and excluded, and the reasons for exclusions.

Data extraction and reporting

We coded and extracted data from the 187 eligible articles according to the following aspects: (i) whether the study performed an *a priori* or *a posteriori* generalizability assessment, or both; (ii) the compared populations and the conclusions of the assessment; (iii) the result outputs (e.g., generalizability scores, descriptive comparison); (iv) the focused disease; (v) the focused population subgroup (e.g., elderly); (vi) the types of the real-world data (RWD) used to profile the target population (i.e., trial data, hospital data, regional data, national data, and international data). Note that trial data can also be regional, national, or even international, depending on the scale of the trial. Regardless, we considered them in the category of "trial data" as the study population of a trial is typically small compared with observational cohorts or RWD. For observational cohorts or RWD (e.g., EHRs), we extracted the scale of the databases (i.e., single hospital, regional, national, and international). For studies that compared characteristics of different populations to indicate generalizability issues, we further coded the populations that were compared (e.g., enrolled patients, eligible patients, general population, ineligible patients), and the types of characteristics that were compared (i.e.,

Table 1 Inclusion and exclusion criteria for articles

Type	Criteria
Inclusion criteria	Articles about generalizability assessment of clinical trial(s) on a specific treatment (e.g., medication, device, or medical procedure)
	Articles must compare the study sample or eligible patients with the patients not in trials
Exclusion criteria	Conference abstracts or nonresearch articles
	Articles about assessing the external validity of screening tools, rating scales, scores, prediction models, etc.
	Articles about the recruitment process of a trial or multiple trials (including certain systematic review articles)
	Articles about the use of eligibility criteria of a trial or multiple trials (including certain systematic review articles)
	Articles about the setting of a trial or multiple trials (e.g., hospital size)
	Articles that promised to consider external validity in future work
	Articles that responded to another article
	Articles that considered outcomes that are not health-related

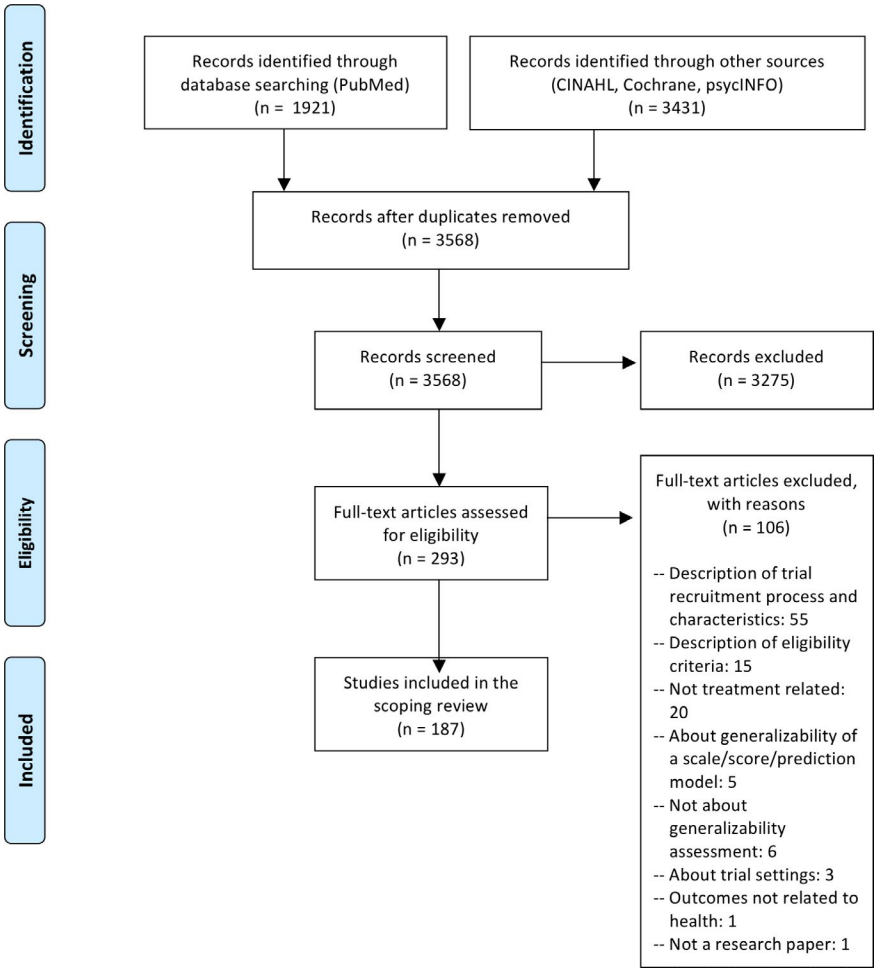

Figure 1 The PRISMA flow diagram of the review. PRISMA, Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

demographic information, clinical attributes and comorbidities, outcomes, and adverse events). We then used Fisher's exact test to assess whether there is a difference in the types of characteristics between *a priori* and *a posteriori* generalizability assessment studies.

INTERPRETATION OF AVAILABLE INFORMATION

Categorization and characteristics of generalizability assessment studies

As shown in **Figure 2**, there was an increasing number of generalizability assessment studies from 1985 to April 2019.

Among the 187 articles, only 14 are methods articles, of which 12 studies have evaluated the proposed methods and applied them to specific clinical trials as examples, whereas the other 2 used simulated data to demonstrate their utility. See the tab "Methods Papers" in **Data Set 1** for details.

Figure 3 shows a taxonomy that synthesizes existing generalizability assessment methods. We defined three major dimensions: (i) time perspective corresponding to data availability; (ii) output (i.e., score vs. nonscore) of the generalizability assessment results; and (iii) populations of interest. **Figure 3a,b** lists the different types of populations being compared in *a priori* and *a posteriori* generalizability assessments, respectively. **Table 2** shows the number of articles along with references of representative articles for each type of the method. Note that "*Post-hoc* generalization" should be considered as a subtype of the *a posteriori* method in which statistical methods were applied to generalize the results of a clinical trial to the broader target population. For example, Westreich *et al.*[18] proposed a method that uses an inverse odds weighting approach to estimate the treatment effect of the trial results in the target population. Complete information about the 187 included articles is shown in **Data Set 1**.

Time perspective of generalizability assessment in terms of data availability

Of the 187 studies, 57 (30.5%) assessed *a priori* generalizability, 109 (58.3%) assessed *a posteriori* generalizability, and 17 (9.1%) assessed both *a priori* and *a posteriori* generalizability. Among the 109 *a posteriori* studies, 17 used propensity scores or other weighting methods to weight the study population while reducing the randomization bias, and then compared the characteristics of the weighted study population with the target population.[19] Four studies fall into the *post hoc* generalization category that investigated how the results can be generalized to the target populations.[18,20–22] **Figure 4** shows the increasing trends

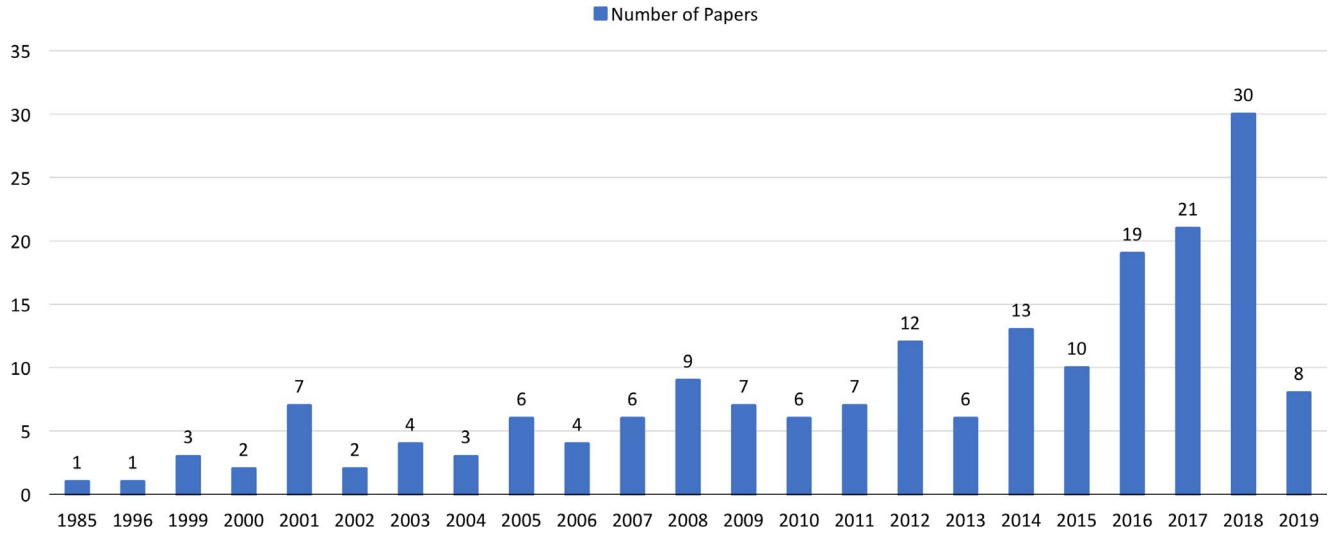

Figure 2 The numbers of generalizability assessment studies from 1985 to April 2019.

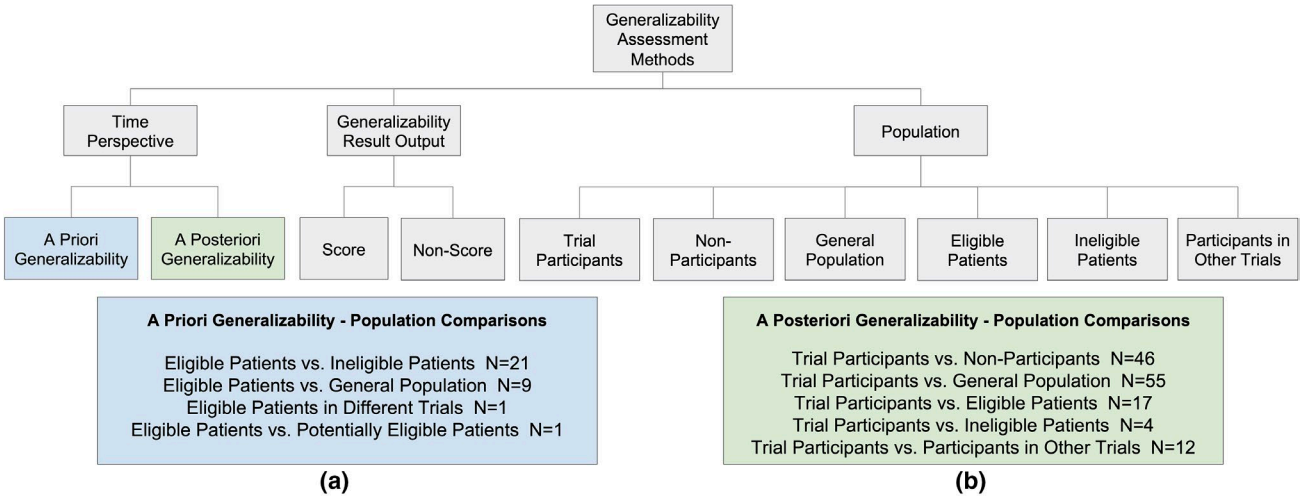

Figure 3 A taxonomy of generalizability assessment methods. Boxes (**a**) and (**b**) list the different types of populations compared in *a priori* and *a posteriori* generalizability assessment articles, respectively.

of both *a priori* and *a posteriori* generalizability assessment studies in the past 30 years. Before 2015, there were slightly more studies on assessing *a posteriori* generalizability than *a priori* generalizability and this difference became more significant after 2015.

Comparisons of populations in generalizability assessment studies

Among the 187 studies, 144 (77.0%) compared the enrolled or eligible patients with observational data collected in routine care. The *a priori* generalizability studies compared eligible patients (by applying the eligibility criteria on a patient database) with: (i) ineligible patients (N = 21); (ii) potentially eligible patients (N = 1); (iii) the general population (N = 9); or (iv) eligible patients in other trials (N = 1). The *a posteriori* generalizability studies compared trial participants with: (i) nonparticipants (those who do not meet exclusion criteria of a trial or those who were eligible for a trial but not randomized) (N = 46); (ii) the general population (N = 55); (iii) eligible patients (N = 17); (iv) ineligible patients (N = 4); or (v) participants in other trials (N = 12). One *a posteriori* generalizability study compared the different participant subgroups in a trial.[23] In general, we excluded studies that merely compared the patients in different arms of a single trial; nevertheless, this study was included as it used broad inclusion and minimal exclusion criteria to evaluate whether phase III clinical trials can recruit representative depressed outpatients.[23] **Table 3** shows the number of generalizability studies by different combinations of compared study-vs.-target population types as well as the types of patient information (e.g., demographics, clinical outcomes) that were compared. Among the 144 studies, 94.4% (N = 136) compared populations' demographics; 81.3% (N = 117) compared clinical characteristics; 44.4% (N = 64) compared treatment outcomes; and very few (4.9%, N = 7) compared adverse events. The result of Fisher's exact test (see

Table 2 Categorization of generalizability assessment methods

Axis	Item	Number of publications (*N* = 187)	Example article
Types of methods	*A priori*	57	Zimmerman *et al.*[13]
	A posteriori	113[a]	Cahan *et al.*[34]
	Post hoc generalization[b]	4	Cole *et al.*[20]
	A priori/a posteriori	17	Lane *et al.*[67]
Output of results	Score	9	Weng *et al.*[25]
	Nonscore	178	Westreich *et al.*[18]

[a]Including the four *post hoc* generalization studies. [b]*Post hoc* generalization: studies that applied methods to generalize a trial's results to the broader target population (e.g., estimate the treatment effect in the target population with the trial results without recruiting and collecting more participant data).

Table S2 in **Supplementary File I**) shows that *a posteriori* generalizability studies were more likely to compare demographic information than *a priori* generalizability studies. With respect to the conclusions about the generalizability of the evaluated trials, 29.4% (*N* = 55) concluded that the trials are generalizable, 59.4% (*N* = 111) concluded that they are not generalizable, and 11.2% (*N* = 21) reported mixed or neutral results in which parts of the analysis indicated good generalizability, whereas the other parts did not.

Output of generalizability assessment results

Only nine studies used a score to quantify the generalizability of a trial or trial set. Among 74 *a priori* generalizability studies, only five analyzed generalizability with score-based methods. Most score-based *a priori* generalizability assessment methods were developed by informaticians.[24] These informatics-based *a priori* methods, such as the Generalizability Index for Study Trait (GIST),[25] mGIST,[26] and GIST 2.0,[27] aimed to quantify the population representativeness of trials using the trial's eligibility criteria combined with the target population's demographic and clinical characteristics corresponding to those criteria. For example, the GIST score quantifies the population representativeness of multiple studies with respect to a single study criterion.[25] It 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. mGIST extended GIST to a multivariate setting by creating combinations of non-overlapping value intervals of multiple study criteria.[26] However, mGIST did not consider the importance of each variable in terms of its restrictiveness for patient selection; thus, GIST 2.0 assigns weights corresponding to variable importance to assess the population representativeness of a trial with respect to either a single study trait (sGIST) or multiple study traits (mGIST 2.0).[27] Previously, Sen et al. have demonstrated the correlation between GIST 2.0 and the adverse events of the patients enrolled in clinical trials[28]. Nevertheless, these methods could be further validated to show the strong correlation between generalizability scores with the outcomes of patients in the target population (e.g., treatment outcomes, adverse events).

Of 74 *a priori* generalizability studies, 69 are non–score-based with two major types: (i) studies that applied a standard set of eligibility criteria representative of clinical trials on a disease and assess how many patients in a database would fulfill typical eligibility criteria[29]; and (ii) studies that descriptively compared the demographic and/or clinical characteristics between eligible patients and a target population (e.g., general population in routine care,[30] and ineligible patients[31]).

There are 122 studies that utilized non–score-based *a posteriori* methods,. For example, Susukida *et al.*[32]

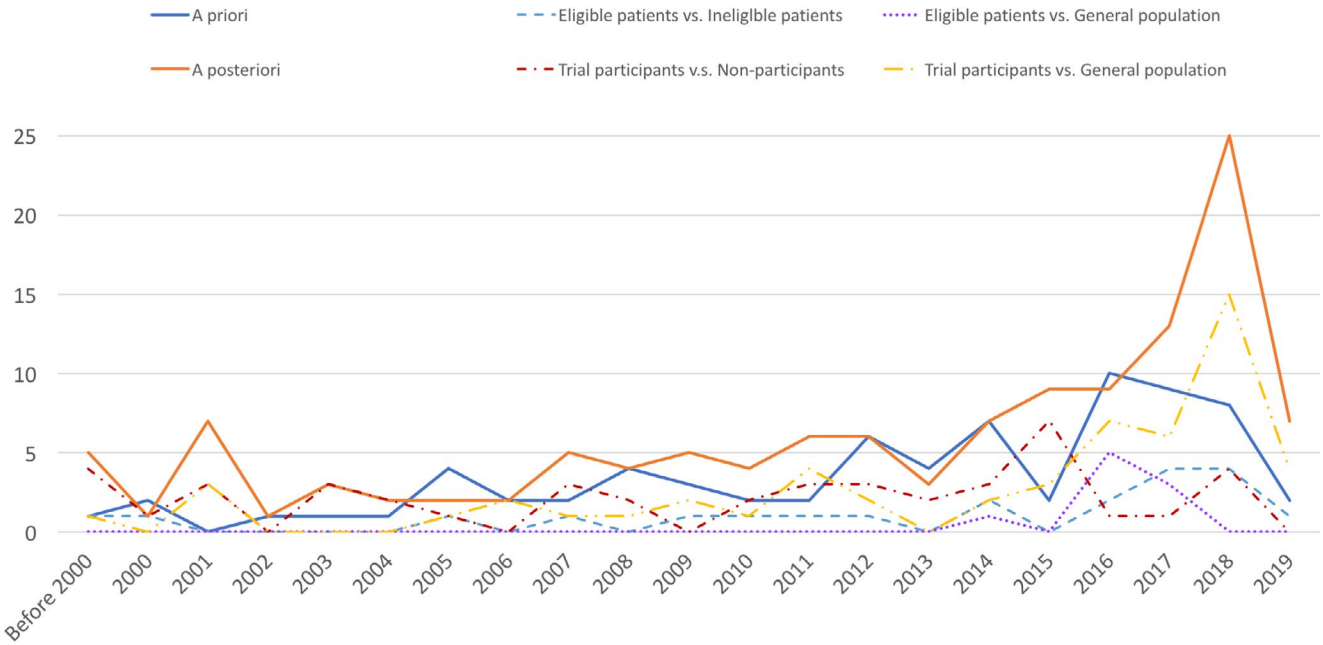

Figure 4 The yearly trend of generalizability assessment publications by methods in terms of data availability.

Table 3 Studies comparing a study population with a target population

Combinations of study population and compared target population	Numbers of articles (*N* = 144)	Compared patient information				Example article
		Demographic information (*N* = 136, 94.4%)	Clinical characteristics (*N* = 117, 81.3%)	Outcomes (*N* = 64, 44.4%)	Adverse events (*N* = 7, 4.9%)	
Trial participants	Nonparticipants (excluded by the trial, or eligible but nonrandomized)	46	37	23	3	Agweyu *et al.*[41]
Trial participants	General population	55	54	42	23	1	McClure *et al.*[68]
Trial participants	Eligible patients (by applying eligibility criteria on the patient data)	17	16	16	6	1	Arora *et al.*[53]
Trial participants	Ineligible patients (by applying criteria on the general population)	4	4	3	2	0	Laskay *et al.*[48]
Trial participants	Participants in other trials	12	12	10	5	1	Laffin *et al.*[69]
A subgroup of trial participants	Trial participants of the same trial but in other subgroups	1	1	1	1	0	Wisniewski *et al.*[23]
Eligible patients	Ineligible patients (by applying criteria on the general population	21	17	17	11	0	Saeed *et al.*[31]
Eligible patients	Potentially eligible patients	1	1	1	0	0	Malatestinic *et al.*[70]
Eligible patients	Eligible patients in other trials	1	0	1	0	0	Fortin *et al.*[71]
Eligible patients	General population	9	9	9	1	1	Weng *et al.*[25]

assessed the difference in the mean propensity scores to compare the differences between the study sample and the target population. Moore *et al.*[33] compared the demographic, clinical, and laboratory characteristics between human immunodeficiency virus (HIV)–infected participants in two antiretroviral trials and eligible patients. The non–score-based *a posteriori* or *a priori* methods that only descriptively compare demographic data between different cohorts lack rigorous validation that associates the measured generalizability with outcomes in the target populations.

Very few (*N* = 4) score-based *a posteriori* methods exist. Cahan *et al.*[34] proposed a framework to produce a "generalizability score" that quantifies the relative difference of a demographic or clinical attribute between the enrolled patients in different trials (i.e., the difference of an attribute is the ratio between the attribute values in the two compared studies). Stuart *et al.*[35] used a propensity-score–based metric to quantify the similarity between the participants in a RCT and the target population. It weights the control group outcomes and assesses how well the propensity-score–adjusted outcomes track the outcomes observed in the target population. Susukida *et al.*[36] used the pooled difference in the mean propensity scores between the RCTs and the target population to quantify the population representativeness of RCTs. **Table S3** in **Supplementary File I** shows these examples with more detailed information about their methods.

Disease areas of generalizability assessment

Generalizability assessments have been conducted on trials of various disease areas, including cancer (*N* = 35; e.g., Sam *et al.*[37]), cardiovascular diseases (*N* = 34; e.g., Patel *et al.*[38]), mental diseases (*N* = 33; e.g., Zimmerman *et al.*[13]), musculoskeletal diseases (*N* = 8; e.g., Becker *et al.*[39]), HIV/acquired immunodeficiency syndrome (*N* = 6; e.g., Saeed *et al.*[31]), endocrine diseases (*N* = 6; e.g., Wittbrodt *et al.*[40]), drug or alcohol abuse (*N* = 6; e.g., Susukida *et al.*[36]), respiratory diseases (*N* = 5; e.g., Agweyu *et al.*[41]), and smoking (*N* = 5; Susukida *et al.*[12]), surgery (*N* = 3; e.g., Fischer *et al.*[42]), ear diseases (*N* = 3; e.g., Rovers *et al.*[43]), digestive disease (*N* = 3; Millard *et al.*[44]), sleep disorders (*N* = 3; Huls *et al.*[45]), skin diseases (*N* = 3; Yiu *et al.*[46]), pain (*N* = 2; de C Williams *et al.*[47]), and other diseases (*N* = 11; e.g., Laskay *et al.*[48]). 21 articles did not specifically focus on a particular disease (e.g., Hong *et al.*[49]).

Data sources used to define target populations

Figure 5 depicts the trends of the different types of data used for profiling the target population in generalizability studies. "Trial-data" are data from patients considered for trials (but not enrolled); "Hospital-data" indicate that the patient data were from small group (i.e., 1–3) of hospitals; and "region-/national-/international-levels" refer to the scale of the hospital/registry/survey data. It is evident that hospital data, national (e.g., Epidemiology, and End Results (SEER) data,[50] National Health and Nutritional Examination Survey,[40] UK Clinical Practice Research Datalink[49]), and international data (e.g., Global Registry of

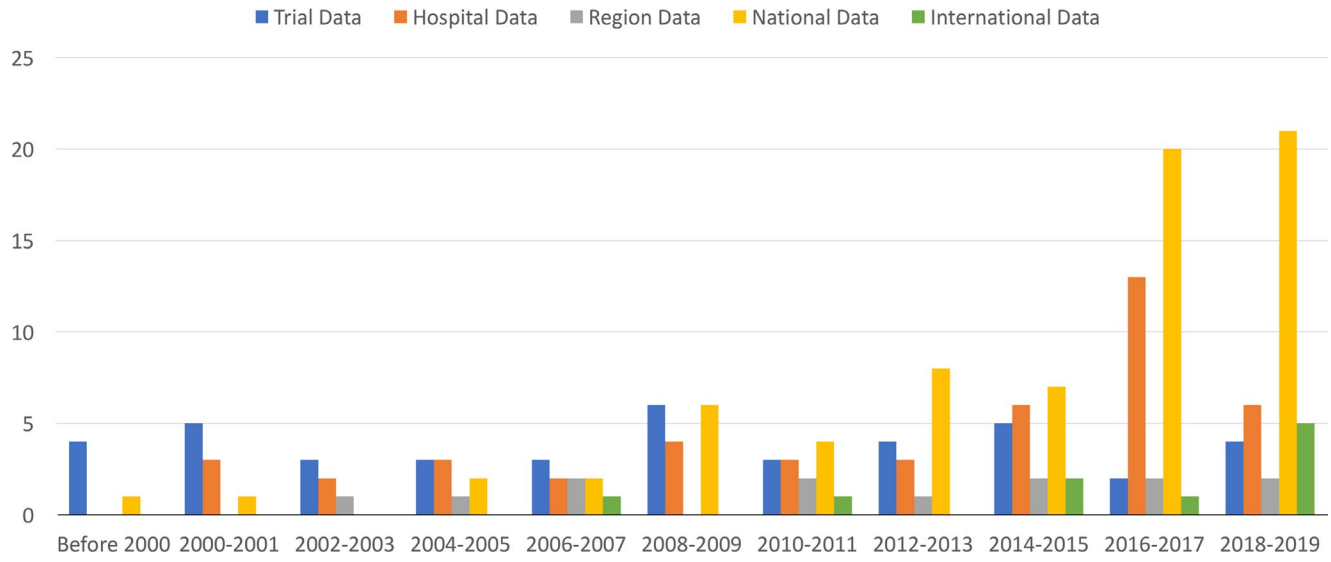

Figure 5 Trends of the data source types used for profiling the target populations.

Acute Coronary Events[51]) have been used more frequently over time.

Focused population subgroups

Of the 187 studies, 28 (15%) studies focused on the underrepresentation of specific population subgroups: children ($N = 8$); elderly ($N = 12$); gender ($N = 9$); and ethnic minorities ($N = 6$). The elderly population is the most studied underrepresented subgroup. Note that some studies discussed more than one subgroup. For example, Heiat *et al.*[52] analyzed the enrolled patients in 59 heart failure clinical trials and found that older adults and female and nonwhite patients were underrepresented in these trials.

IMPLICATIONS AND FUTURE DIRECTIONS

Over the past 2 decades, an increasing number of studies have assessed the generalizability of clinical trials, especially after 2015. Although the literature on generalizability assessment and associated methods is abundant, our review has been shown that it is poorly organized and there is little agreement on analytic procedures.

Among the studies we reviewed, most generalizability assessments were conducted *a posteriori* rather than *a priori*, hence could only discover generalizability issues after the completion of a trial, missing the opportunity for early detection and correction of sampling procedures. In addition, we found that most generalizability assessments are shallow: (i) in *a priori* generalizability studies, researchers often apply the study eligibility criteria on a patient database (e.g., EHRs from a hospital) to identify the study population and compare patient demographics, clinical characteristics, and outcomes between the study population and a target population; and (ii) in *a posteriori* generalizability studies, researchers make comparisons of different types of patient characteristics between the enrolled patients and a target population. In a few studies,[46] researchers first used the propensity score or other weighting mechanisms to reduce the bias of randomization of patients into intervention arms or control arms and then compared the weighted study population with the target population. We also observed that, for the 144 studies (see **Table 3**) that compared enrolled patients or eligible patients with observational data collected in routine care, only 7 (4.9%) compared the adverse events between these populations, leaving an important gap to fill in future generalizability assessments.

Score-based generalizability assessment methods are scarce in both *a priori* ($N = 5$) and *a posteriori* ($N = 4$) studies, representing a lost opportunity to quantify a study's generalizability. For example, a score-based *a priori* generalizability method can yield actionable knowledge to help investigators adjust the eligibility criteria toward improved population representativeness (i.e., a higher generalizability score), while balancing the trial's internal validity, before the trial starts enrollment.

Not surprisingly, we observed that there is no universal definition of the "target population," due in part to the evolving nature of treatment development (e.g., drug repurposing), but also to the lack of consensus on the applicability of a trial. In fact, specifying the target population is difficult not only in generalizability assessment but also in clinical practice. Regulatory agencies (e.g., the US Food and Drug Administration (FDA)) typically only approve a treatment agent with an indication that its use is restricted to the study population tested in the trials; nonetheless, "off-label" use of the agent is very common. Because it is virtually impossible to assess the data for all potential patients in the target population, generalizability assessment studies mostly use a convenience sample (e.g., patients with a specific condition in an observational database) to approximate the target population. Traditionally, researchers compare characteristics between the enrolled patients with the eligible but nonrandomized patients,[53] so they are limited to studying patients who are geographically close to the study site. In recent years, we observed an increasing trend toward using large-scale, national and international data sets to identify

the target population when assessing study generalizability. With the wide adoption of EHR systems, secondary use of hospital data has increased tremendously.[54] With more observational real-world data (e.g., data from the Patient-Centered Clinical Research Network (PCORnet)[55]) becoming readily available, we anticipate that both *a priori* and *a posteriori* generalizability assessment will become *de facto* processes in trial design and conduct.

In this review we also found that no study has investigated the trade-off between clinical trial generalizability (external validity) and internal validity. As this is a critical problem in clinical research, we hope that this work can encourage the research community to design novel approaches to afford balance to this issue. Such work may need to account for study-specific methodology as well as the primary end point of the trial. Internal validity may be a higher priority than generalizability in early-phase studies where determination of dose-limiting toxicities is the primary objective.

Importance of *a priori* generalizability assessment in eligibility criteria design process

Conventionally, the eligibility criteria design of a trial depends on investigators' empirical knowledge of the disease, drug, and the trial. Frequently, criteria are adopted from previous similar protocols without due consideration of the differing drug effects or patient populations,[3] leading to propagation of difficult-to-justify criteria.[56] Van Spall *et al.*[57] reviewed 283 RCTs between 1994 and 2006 and reported that 37% of the trials' eligibility criteria were poorly justified, and 84% of the trials had at least one poorly justified exclusion criterion. Poorly justified and unnecessarily restrictive criteria limit patients' access to trials and lead to low study accrual rates,[58] resulting in studies that fail to be completed[59] or fail to capture the heterogeneity of the target population (e.g., leading to unintended serious adverse events after the approval of the treatments[3]). In particular, people aged ≥ 65 years are still significantly underrepresented in drug trials, especially cancer trials.[60] Conducting *a priori* generalizability assessment during trial design can be beneficial because eligibility criteria can then be appropriately and objectively adjusted (i.e., with the *a priori* generalizability score) to include a diverse population in the trial before it is conducted.

Nevertheless, there are a number of barriers to adopting *a priori* generalizability assessments, such as: (i) although some informatics-based methods such as GIST 2.0,[27] have been validated against adverse events extracted from results of clinical trial enrolled patients[28], we think it is important to further validate them against real-world patient outcomes and adverse events in the target populations; (ii) the lack of readily available, well-vetted statistical and informatics tools; and (iii) the knowledge gap in best practice for generalizability assessment. Further, there is a tacit belief that traditional standards—making eligibility criteria unnecessarily restrictive—need to be maintained for fear of exposing trial patients to harm and rejection by regulatory and safety monitoring bodies.[5] Thus, trial investigators do not necessarily feel empowered to modify these criteria in the absence of data or a directive to do so.

Informatics' opportunities for *a priori* generalizability assessments

Streamlining *a priori* generalizability assessment requires automated cohort discovery from RWD, such as EHRs. Recently, significant national efforts have started building tools and algorithms to support cohort discovery for clinical trials. For example, the i2b2 (Informatics for Integrating Biology and the Bedside)[61] cohort discovery tool is widely deployed and used, and the CALYPSO[62] tool based on the OMOP (Observational Medical Outcome Partnership) Common Data Model (CDM) is also emerging. Nevertheless, these tools require investigators to manually translate eligibility criteria into cohort discovery queries, posing a significant barrier. Automated generalizability assessment requires computable phenotypes.[63] With a computable eligibility criteria (CEC) infrastructure,[64] the study population of a trial can be readily identified and compared with the target population.

Making eligibility criteria computable is nontrivial. One approach is to parse free-text eligibility criteria using advanced natural language processing (NLP) methods and then transforming them into executable database queries. For example, Critera2Query was developed to transform free-text criteria into OMOP CDM-based database queries.[65] However, the complexity of eligibility criteria makes it difficult for NLP to achieve optimal results. The performance of two important NLP tasks—entity recognition and relation extraction—in Criteria2Query is suboptimal (i.e., an F1 score of 0.795 and 0.805, respectively).[65] A second approach, including our own prior work,[64] has connected eligibility criteria to underlying clinical databases via ontologies and made them computable through ontology-based data access frameworks. Use of ontology creates a shared, controlled vocabulary of eligibility criteria and standardizes the definitions of data elements, making data understandable to both humans and computers. Although parsing eligibility criteria and standardizing study traits is still largely a manual process in this exploratory phase, it yields much better quality in terms of accuracy in representing eligibility criteria as well as better performance in terms of precision and recall in retrieving cohorts accurately. Nevertheless, as NLP methods advance, there are opportunities to adapt NLP techniques to automate the process to make it more scalable or employ a hybrid approach that increases both accuracy and scalability.

Rather than parsing free-text eligibility criteria after the fact, adopting a CEC-based criteria authoring tool during the trial design phase may be more efficient. Equipped with CEC and readily accessible large, real-world data sets, the tool could be developed to assist trial designs by providing real-time cohort discovery and *a priori* generalizability assessment services. As such, eligibility criteria can be fine-tuned and adequately adjusted to improve trial generalizability during the design phase.

In this review, we have found that existing informatics-based generalizability assessment methods such as GIST,[25] mGIST,[26] and GIST 2.0[27] should be further validated. Their correlations with patient outcomes in real-world populations should be systematically evaluated by informaticians. In addition, an open-source, publicly available toolbox with clear documentation and a guideline should be developed to

aid researchers in choosing appropriate methods to assess their studies' generalizability.

In conclusion, we have systematically organized generalizability assessment methods in a taxonomy consisting of three dimensions: (i) data availability (*a priori* vs. *a posteriori*); (ii) results output (score vs. nonscore); and (iii) populations (e.g., enrolled patients, eligible patients). We observed an increasing trend of generalizability assessment of clinical trials over the past 3 decades. With the wide adoption of EHR systems in the past few years, large-scale, real-world patient data are becoming increasingly promoted (e.g., the FDA's recent effort on the use of real-word data[66]) and available, making generalizability assessment of trials more feasible than ever. However, software tools and packages are still lacking and are not readily available for generalizability assessment. Further, as *a priori* generalizability can be assessed using only study design information (primarily eligibility criteria), it gives investigators a golden opportunity to adjust the study design before the trial starts. Nevertheless, < 40% of studies in our review assessed *a priori* generalizability. Research culture and regulatory policy adaptation are also needed to change the practice of trial design (e.g., relaxing restrictive eligibility criteria) toward better trial generalizability.

Supporting Information. Supplementary information accompanies this paper on the *Clinical and Translational Science* website (www.cts-journal.com).

Funding. This study was supported primarily by the National Institute on Aging of the National Institutes of Health (NIH) under Award Number R21AG061431; and in part by NIH Award UL1TR001427. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflict of Interest. The authors declared no competing interests for this work.

Data Availability Statement. The Excel spreadsheet with all the coded data of the 187 included papers has been submitted with the manuscript (**Data Set 1**). This file has also been deposited to datadryad.org.

1. From the NIH Director: the importance of clinical trials <http://www.nlm.nih.gov/medlineplus/magazine/issues/summer11/articles/summer11pg2-3.html>. Accessed April 1, 2019.
2. Sedgwick, P. External and internal validity in clinical trials. *BMJ.* **344**, e1004 (2012).
3. Beaver, J.A., Ison, G. & Pazdur, R. Reevaluating eligibility criteria—balancing patient protection and participation in oncology trials. *N. Engl. J. Med.* **376**, 1504–1505 (2017).
4. Shields, K.E. & Lyerly, A.D. Exclusion of pregnant women from industry-sponsored clinical trials. *Obstet. Gynecol.* **122**, 1077–1081 (2013).
5. US Food and Drug Administration. Enhancing the diversity of clinical trial populations—eligibility criteria, enrollment practices, and trial designs guidance for industry <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enhancing-diversity-clinical-trial-populations-eligibility-criteria-enrollment-practices-and-trial> (2019). Accessed August 5, 2019.
6. de Jonghe, A., van de Glind, E.M.M., van Munster, B.C. & de Rooij, S.E. Underrepresentation of patients with pre-existing cognitive impairment in pharmaceutical trials on prophylactic or therapeutic treatments for delirium: a systematic review. *J. Psychosom. Res.* **76**, 193–199 (2014).
7. Frank, C. *et al.* Era of faster FDA drug approval has also seen increased black-box warnings and market withdrawals. *Health Aff (Millwood)* **33**, 1453–1459 (2014).
8. Bonell, C., Oakley, A., Hargreaves, J., Strange, V. & Rees, R. Assessment of generalisability in trials of health interventions: suggested framework and systematic review. *BMJ.* **333**, 346–349 (2006).
9. Hoertel, N., LeStrat, Y., Lavaud, P., Dubertret, C. & Limosin, F. Generalizability of clinical trial results for bipolar disorder to community samples. *J. Clin. Psychiatry* **74**, 265–270 (2013).
10. Leiter, A., Diefenbach, M.A., Doucette, J., Oh, W.K. & Galsky, M.D. Clinical trial awareness: changes over time and sociodemographic disparities. *Clin. Trials* **12**, 215–223 (2015).
11. Kanarek, N.F., Kanarek, M.S., Olatoye, D. & Carducci, M.A. Removing barriers to participation in clinical trials, a conceptual framework and retrospective chart review study. *Trials* **13**, 237 (2012).
12. Susukida, R., Crum, R.M., Hong, H., Stuart, E.A. & Mojtabai, R. Comparing pharmacological treatments for cocaine dependence: incorporation of methods for enhancing generalizability in meta-analytic studies. *Int. J. Methods Psychiatry Res.* **27**, e1609 (2018).
13. Zimmerman, M., Walsh, E., Chelminski, I. & Dalrymple, K. Has the symptom severity inclusion requirement narrowed the definition of major depressive disorder in antidepressant efficacy trials? *J. Affect. Disord.* **211**, 60–64 (2017).
14. Kennedy-Martin, T., Curtis, S., Faries, D., Robinson, S. & Johnston, J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* **16**, 495 (2015).
15. Finding what works in health care: standards for systematic reviews <http://www.nationalacademies.org/hmd/Reports/2011/Finding-What-Works-in-Health-Care-Standards-for-Systematic-Reviews.aspx>. Accessed May 1, 2019.
16. Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G.& PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ.* **339**, b2535 (2009).
17. McHugh, M.L. Interrater reliability: the kappa statistic. *Biochem. Med. (Zagreb)* **22**, 276–282 (2012).
18. Westreich, D., Edwards, J.K., Lesko, C.R., Stuart, E. & Cole, S.R. Transportability of trial results using inverse odds of sampling weights. *Am. J. Epidemiol.* **186**, 1010–1014 (2017).
19. Cole, S.R. & Stuart, E.A. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 Trial. *Am. J. Epidemiol.* **172**, 107–115 (2010).
20. Cole, S.R. & Stuart, E.A. Generalizing evidence from randomized clinical trials to target populations: the ACTG 320 trial. *Am. J. Epidemiol.* **172**, 107–115 (2010).
21. Stuart, E.A. & Rhodes, A. Generalizing treatment effect estimates from sample to population: a case study in the difficulties of finding sufficient data. *Eval. Rev.* **41**, 357–388 (2017).
22. Stuart, E.A., Bradshaw, C.P. & Leaf, P.J. Assessing the generalizability of randomized trial results to target populations. *Prev. Sci.* **16**, 475–485 (2015).
23. Wisniewski, S.R. *et al.* Can phase III trial results of antidepressant medications be generalized to clinical practice? A STAR*D report. *Am. J. Psychiatry* **166**, 599–607 (2009).
24. Weng, C. Optimizing clinical research participant selection with informatics. *Trends Pharmacol. Sci.* **36**, 706–709 (2015).
25. Weng, C. *et al.* A distribution-based method for assessing the differences between clinical trial target populations and patient populations in electronic health records. *Appl. Clin. Inform.* **5**, 463–479 (2014).
26. He, Z. *et al.* Multivariate analysis of the population representativeness of related clinical studies. *J. Biomed. Inform.* **60**, 66–76 (2016).
27. Sen, A. *et al.* GIST 2.0: a scalable multi-trait metric for quantifying population representativeness of individual clinical studies. *J. Biomed. Inform.* **63**, 325–336 (2016).
28. Sen, A., *et al.* Correlating eligibility criteria generalizability and adverse events using Big Data for patients and clinical trials. *Ann. N. Y. Acad. Sci.* **1387**, 34–43 (2017).
29. Hoertel, N., Le Strat, Y., Blanco, C., Lavaud, P. & Dubertret, C. Generalizability of clinical trial results for generalized anxiety disorder to community samples. *Depress. Anxiety* **29**, 614–620 (2012).
30. Bress, A.P. *et al.* Generalizability of SPRINT Results to the U.S. adult population. *J. Am. Coll. Cardiol.* **67**, 463–472 (2016).
31. Saeed, S. *et al.* How generalizable are the results from trials of direct antiviral agents to people coinfected with HIV/HCV in the real world? *Clin. Infect. Dis.* **62**, 919–926 (2016).
32. Susukida, R., Crum, R.M., Stuart, E.A. & Mojtabai, R. Generalizability of the findings from a randomized controlled trial of a web-based substance use disorder intervention. *Am. J. Addict.* **27**, 231–237 (2018).
33. Moore, D.A. *et al.* How generalizable are the results of large randomized controlled trials of antiretroviral therapy? *HIV Med.* **1**, 149–154 (2000).
34. Cahan, A., Cahan, S. & Cimino, J.J. Computer-aided assessment of the generalizability of clinical trial results. *Int. J. Med. Inform.* **99**, 60–66 (2017).

35. Stuart, E.A., Cole, S.R., Bradshaw, C.P. & Leaf, P.J. The use of propensity scores to assess the generalizability of results from randomized trials. *J. R. Stat. Soc. Ser. A Stat. Soc.* **174**, 369–386 (2001).
36. Susukida, R., Crum, R.M., Stuart, E.A., Ebnesajjad, C. & Mojtabai, R. Assessing sample representativeness in randomized controlled trials: application to the National Institute of Drug Abuse Clinical Trials Network. *Addiction* **111**, 1226–1234 (2016).
37. Sam, D., Gresham, G., Abdel-Rahman, O. & Cheung, W.Y. Generalizability of clinical trials of advanced melanoma in the real-world, population-based setting. *Med. Oncol.* **35**, 110 (2018).
38. Patel, H.C. *et al.* Assessing the eligibility criteria in Phase III randomized controlled trials of drug therapy in heart failure with preserved ejection fraction: the critical play-off between a "pure" patient phenotype and the generalizability of trial findings. *J. Card. Fail.* **23**, 517–524 (2017).
39. Becker, H.J. *et al.* A novel use of the Spine Tango registry to evaluate selection bias in patient recruitment into clinical studies: an analysis of patients participating in the Lumbar Spinal Stenosis Outcome Study (LSOS). *Eur. Spine J.* **26**, 441–449 (2017).
40. Wittbrodt, E.T. *et al.* Generalizability of glucagon-like peptide-1 receptor agonist cardiovascular outcome trials enrollment criteria to the US type 2 diabetes population. *Am. J. Manag. Care* **24**, S146–S155 (2018).
41. Agweyu, A. *et al.* Comparable outcomes among trial and nontrial participants in a clinical trial of antibiotics for childhood pneumonia: a retrospective cohort study. *J. Clin. Epidemiol.* **94**, 1–7 (2018).
42. Fischer, L. *et al.* To whom do the results of the multicenter, randomized, controlled INSECT trial (ISRCTN 24023541) apply?—assessment of external validity. *BMC Surg.* **12**, 2 (2012).
43. Rovers, M.M., Zielhuis, G.A., Bennett, K. & Haggard, M. Generalisability of clinical trials in otitis media with effusion. *Int. J. Pediatr. Otorhinolaryngol.* **60**, 29–40 (2001).
44. Millard, J.D. *et al.* Assessing the external validity of a randomized controlled trial of anthelminthics in mothers and their children in Entebbe, Uganda. *Trials* **15**, 310 (2014).
45. Huls, H. *et al.* Inclusion and exclusion criteria of clinical trials for insomnia. *J. Clin. Med.* **7**, 206 (2018). https://doi.org/10.3390/jcm7080206
46. Yiu, Z.Z.N. *et al.* A standardisation approach to compare treatment safety and effectiveness outcomes between clinical trials and real world populations in psoriasis. *Br J. Dermatol.* **181**, 1265–1271 (2019).
47. de C Williams, A.C., Nicholas, M.K., Richardson, P.H., Pither, C.E. & Fernandes, J. Generalizing from a controlled trial: the effects of patient preference versus randomization on the outcome of inpatient versus outpatient chronic pain management. *Pain* **83**, 57–65 (1999).
48. Laskay, N.M.B. *et al.* A comparison of the MOMS trial results to a contemporaneous, single-institution, postnatal closure cohort. *Childs Nerv. Syst.* **33**, 639–646 (2017).
49. Hong, J.L. *et al.* Generalizing randomized clinical trial results: implementation and challenges related to missing data in the target population. *Am. J. Epidemiol.* **187**, 817–827 (2018).
50. Costa, L.J., Hari, P.N. & Kumar, S.K. Differences between unselected patients and participants in multiple myeloma clinical trials in US: a threat to external validity. *Leuk. Lymphoma* **57**, 2827–2832 (2016).
51. Steg, P.G. *et al.* External validity of clinical trials in acute myocardial infarction. *Arch. Intern. Med.* **167**, 68–73 (2007).
52. Heiat, A., Gross, C.P. & Krumholz, H.M. Representation of the elderly, women, and minorities in heart failure clinical trials. *Arch. Intern. Med.* **162**, 1682–1688 (2002).
53. Arora, S. *et al.* Cytoreductive nephrectomy: assessing the generalizability of the CARMENA trial to real-world national cancer data base cases. *Eur. Urol.* **75**, 352–353 (2019).
54. Botsis, T., Hartvigsen, G., Chen, F. & Weng, C. Secondary use of EHR: data quality issues and informatics opportunities. *Summit Transl. Bioinform.* **2010**, 1–5 (2010).
55. PCORnet, the National Patient-Centered Clinical Research Network <http://www.pcornet.org/>. Accessed May 1, 2019.
56. Hao, T., Rusanov, A., Boland, M.R. & Weng, C. Clustering clinical trials with similar eligibility criteria features. *J. Biomed. Informat.* **52**, 112–120 (2014).
57. Van Spall, H.G., Toren, A., Kiss, A. & Fowler, R.A. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. *JAMA* **297**, 1233–1240 (2007).
58. Lemieux, J. *et al.* Identification of cancer care and protocol characteristics associated with recruitment in breast cancer clinical trials. *J. Clin. Oncol.* **26**, 4458–4465 (2008).
59. Gerber, D.E., Pruitt, S.L. & Halm, E.A. Should criteria for inclusion in cancer clinical trials be expanded? *J. Comp. Eff. Res.* **4**, 289–291 (2015).
60. Abbasi, J. Older patients (still) left out of cancer clinical trials. *JAMA* **322**, 1751–1753 (2019). https://doi.org/10.1001/jama.2019.17016.
61. i2b2 (Informatics for Integrating Biology and the Bedside) <https://www.i2b2.org/>. Accessed May 1, 2019.
62. CALYPSO: criteria assessment logic for your population studies of observations <http://www.ohdsi.org/web/calypso/#/>. Accessed May 1, 2019.
63. Wiese, A.D. *et al.* Performance of a computable phenotype for identification of patients with diabetes within PCORnet: the Patient-Centered Clinical Research Network. *Pharmacoepidemiol. Drug Saf.* **28**, 632–639 (2019).
64. Zhang, H. *et al.* Computable eligibility criteria through ontology-driven data access: a case study of hepatitis C virus trials. *AMIA Ann. Symp. Proc.* **2018**, 1601–1610 (2018).
65. Yuan, C. *et al.* Criteria2Query: a natural language interface to clinical databases for cohort definition. *J. Am. Med. Inform. Assoc.* **26**, 294–305 (2019).
66. Sherman, R.E., Davies, K.M., Robb, M.A., Hunter, N.L. & Califf, R.M. Accelerating development of scientific evidence for medical products within the existing US regulatory framework. *Nat. Rev. Drug Discov.* **16**, 297–298 (2017).
67. Lane, K. *et al.* African American screening and enrollment in (clot lysis: evaluating accelerated resolution of intraventricular hemorrhage III) CLEAR III. *Clin. Res. (Alex.)* **32**, (2018).
68. McClure, E.A. *et al.* Comparing adult cannabis treatment-seekers enrolled in a clinical trial with national samples of cannabis users in the United States. *Drug Alcohol Depend.* **176**, 14–20 (2017).
69. Laffin, L.J., Besser, S.A. & Alenghat, F.J. A data-zone scoring system to assess the generalizability of clinical trial results to individual patients. *Eur. J. Prev. Cardiol.* **26**, 569–575 (2019).
70. Malatestinic, W. *et al.* Characteristics and medication use of psoriasis patients who may or may not qualify for randomized controlled trials. *J. Manag. Care Spec. Pharm.* **23**, 370–381 (2017).
71. Fortin, M. *et al.* Randomized controlled trials: do they have external validity for patients with multiple comorbidities? *Ann. Fam. Med.* **4**, 104–108 (2006).