JAMIA Open, 4(3), 2021, 1–11
doi: 10.1093/jamiaopen/ooab077
Research and Applications

Research and Applications

Quantifying representativeness in randomized clinical trials using machine learning fairness metrics

Miao Qi[1], Owen Cahan[2], Morgan A. Foreman[3], Daniel M. Gruen[2], Amar K. Das[3], and Kristin P. Bennett[1,2]

[1]Department of Computer Science, Rensselaer Polytechnic Institute, Troy, New York, USA, [2]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, New York, USA, and [3]Center for Computational Health, IBM Research, Cambridge, Massachusetts, USA

Corresponding Author: Kristin P. Bennett, Department of Mathematical Sciences, PhD, Rensselaer Polytechnic Institute, 110 8th Street, CII 3129, Troy, NY 12180, USA; bennek@rpi.edu

Received 18 June 2021; Revised 19 August 2021; Editorial Decision 25 August 2021; Accepted 3 September 2021

ABSTRACT

Objective: We help identify subpopulations underrepresented in randomized clinical trials (RCTs) cohorts with respect to national, community-based or health system target populations by formulating population representativeness of RCTs as a machine learning (ML) fairness problem, deriving new representation metrics, and deploying them in easy-to-understand interactive visualization tools.
Materials and Methods: We represent RCT cohort enrollment as random binary classification fairness problems, and then show how ML fairness metrics based on enrollment fraction can be efficiently calculated using easily computed rates of subpopulations in RCT cohorts and target populations. We propose standardized versions of these metrics and deploy them in an interactive tool to analyze 3 RCTs with respect to type 2 diabetes and hypertension target populations in the National Health and Nutrition Examination Survey.
Results: We demonstrate how the proposed metrics and associated statistics enable users to rapidly examine representativeness of all subpopulations in the RCT defined by a set of categorical traits (eg, gender, race, ethnicity, smoking status, and blood pressure) with respect to target populations.
Discussion: The normalized metrics provide an intuitive standardized scale for evaluating representation across subgroups, which may have vastly different enrollment fractions and rates in RCT study cohorts. The metrics are beneficial complements to other approaches (eg, enrollment fractions) used to identify generalizability and health equity of RCTs.
Conclusion: By quantifying the gaps between RCT and target populations, the proposed methods can support generalizability evaluation of existing RCT cohorts. The interactive visualization tool can be readily applied to identified underrepresented subgroups with respect to any desired source or target populations.

Key words: population representativeness, machine learning, randomized clinical trials, subgroup, health equity

© The Author(s) 2021. Published by Oxford University Press on behalf of the American Medical Informatics Association.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (https://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

LAY SUMMARY

Inequitable representation of minority groups and diverse subpopulations in randomized clinical trials (RCTs) can contribute to unfair and avoidable differences in population health outcomes. Standardized methods are needed to assess potential representation disparities between RCT cohorts and the broader populations who could benefit from novel interventions. We show how machine learning fairness metrics used in artificial intelligence applications can be adapted to create metrics that quantify the representativeness of clinical trial cohorts with respect to desired trait-specific subgroups. We demonstrate the scalable representativeness metrics by comparing subgroups in 3 landmark RCTs in diabetes and heart disease with corresponding prevalence in national US population. Supplementary visualizations and statistical tests built on our proposed metrics allow a diversity of researchers from different fields such as computer scientists, clinical researchers, and physicians, to rapidly discover and assess potential disparities in representation of subgroups. Our approach enables users to determine underrepresentation, absence, or overrepresentation of subgroups indicating potential limitations of RCTs. Here, we consider a *posteriori* evaluation of applicability of RCT results to a target population, but the method could be extended to design of new RCTs, and monitoring of RCT enrollment in the future.

BACKGROUND AND SIGNIFICANCE

Inequitable representation and evaluation of diverse subgroups in randomized clinical trials (RCTs) and other clinical research may generate unfair and avoidable differences in population health outcomes.[1–4] In an analysis of trials conducted by Pfizer between 2011 and 2020, scientists found an urgent need for solutions to enhance diverse representation across all populations within clinical research.[5] Similarly, health inequity attracted great public attention during the COVID-19 pandemic.[6–8] For example, race and ethnicity are identified factors associated with risk for COVID-19 infection and mortality.[9–11] Representative enrollment of participants with diverse race and ethnicity is required in clinical trials to ensure valid treatment effect conclusions and to support reliable generalizability of clinical trial results across subpopulations.

A well-designed RCT is considered the most reliable way to estimate cause–effect relationships between treatments and outcomes.[12,13] The randomization process, which makes RCTs gold standards of treatment effectiveness, contains 2 randomization processes, the random sampling from source population to trial cohort and the random assignment from trial cohort to different experimental groups.[14,15] The random sampling is critical to the applicability and generalizability of clinical findings[16–18] but has received much less attention than random assignment. Figure 1 demonstrates that if a latent patient trait guides the patient enrollment into the study and affects the outcome, then the study generalizability to other reference populations may be limited from a causal inference perspective.

Population representativeness and previous works

We define RCT representativeness as the similarity between an RCT cohort and an investigator-defined target population with the specific goal of understanding the representation differences within subpopulations. The target population for an RCT may be different from the population of all individuals who have a particular health condition. For example, Pradhan et al[19] reported that the level of trial representativeness changes if the target population shifts from patients with type 2 diabetes who are eligible to receive liraglutide to all patients with type 2 diabetes, since the potential subjects become younger and are less likely to have comorbidities. Thus, the first step is to let investigators define the target population based on an appropriate real-world data source, such as an Electronic Health Record (EHR) system, or a nationally representative population sample, such as the National Health and Nutrition Examination Survey (NHANES).

Our goal is to calculate representation metrics for all possible subgroups created by the multiple traits and then focus on visualizations and statistical methods that enable users to effectively identify significantly underrepresented subgroups with respect to the target population. Our work complements currently available measurements of trial representativeness. For instance, sGIST, mGIST,

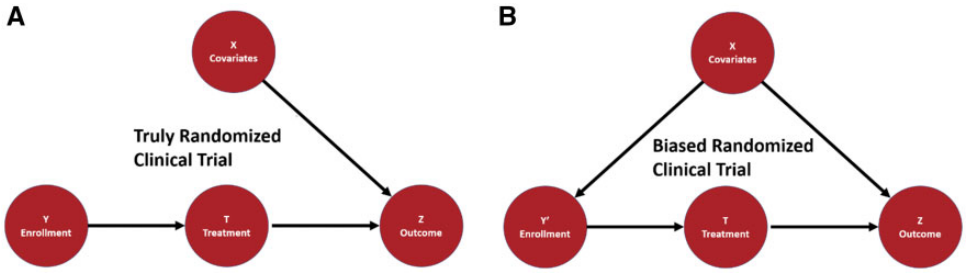

Figure 1. The causal models of truly randomized clinical trials and biased randomized clinical trials. *X* represents the subject covariates; *Y* is the sampling of a subject to a trial; *T* indicates the treatment; and *Z* is the outcome. The black arrows represent causal dependencies between variables. **A.** In the causal model for truly randomized clinical trials, no dependency should exist between *X* and *Y*. Thus, the observed probability of outcome *Z* given the treatment is a good estimate of whether the treatment causes the outcome. **B.** In the causal model for biased randomized clinical trials, an arrow exists between *X* and *Y'*, which indicates the dependence. Thus, invalid causal inferences may be estimated for treatment efficacy among some subpopulations and result in unfair and avoidable population health disparities.

GIST, and GIST 2.0 are a series of *a priori* generalizability method that can calculate generalizability scores on multiple traits across multiple clinical trials with explicitly consideration on eligibility criteria dependences.[20–23] These metrics help researchers identify underrepresented subgroups due to eligibility requirements and can thus be used to inform eligibility criteria in trial design. Our current tool focuses on *a posteriori* evaluation of representativeness, and the analyses presented here deal with simple eligibility criteria such as age over 50 or without diabetes. Other complex eligibility criteria including trait dependencies are left as future work to incorporate GIST 2.0 into our framework.

Our general methodology for calculating and visualizing subgroup representativeness and their statistical significance could also be combined with existing methods for comparing characteristic distributions between study samples and target populations.[24–27] These include basic metrics such as the difference or ratio of subgroup proportions in RCT cohorts and target populations and propensity score methods. These basic metrics are important indicators of population representativeness. But extending them to handle high-dimensional data or small-size subgroups can be challenging since the misrepresentation may be statistically insignificant and hard to detect in visualizations.

Our multi-faceted assessment framework to evaluate diversity, inclusion, and equity provides a comprehensive and interpretable subpopulation-level understanding of population representativeness of RCTs. Our *a posteriori* metrics have defined a significant difference threshold and equity thresholds supported by well-developed guidelines. We show that sunburst visualizations can explicitly present the influence of different variables over the others thus adding more valuable insights to the approach. By indicating the representativeness of all possible subgroups, our approach could eventually help illuminate the "black box" of sample selection and trial generalizability in clinical trials.

Machine learning fairness and previous works

Machine learning (ML) fairness metrics have been developed to quantify and mitigate bias in ML and artificial intelligence (AI) models.[28–30] To improve the performance of existing RCT representativeness measurements, we consider sampling to the RCT a random binary classification problem and develop standardized metrics for RCTs based on variations of ML fairness metrics by mapping to the context of RCTs. ML fairness metrics quantify potential bias toward protected groups in trained ML classification model outcomes. Our metrics, instead of comparing positive and negative classes based on model outcomes, focus on the trial-subject data generation process within the RCT. Our novel insight is to regard subject sampling to an RCT as a classification function that is random and then create variants of ML fairness metrics.

Our metrics capture how well the actual enrollment of subjects to an RCT cohort matches a truly random sampling. The statistical properties of the hypothetical random sampling from a target population can be estimated using nationally representative datasets or clinical databases of individual characteristics, such as NHANES[31] or from EHRs.

Consolidated standards of reporting trials and previous works

Our main goal is to identify all subgroups that are not well represented in RCT study samples in order to understand generalizability with respect to a target population. Our method augments the Consolidated Standards of Reporting Trials (CONSORT)[32,33] statement and its extension CONSORT-Equity,[34] which aims to avoid biased results from incomplete or nontransparent research reports that could mislead decision-making in healthcare. By appropriately defining the target population (such as all individuals with the health condition or those clinically defined by eligibility criteria), our metrics and visualization can support incorporating representativeness evaluation before, during, and after any RCTs. Additionally, they can help an Institutional Review Board (IRB) or funding agency evaluate the equity in trial-design stages and assist government regulators to ensure a fair distribution of clinical benefits from a study to the general population.

Our proposed representativeness metrics are expected to identify subgroups that are insufficiently recruited into and represented in the clinical trial cohort using study summary data only, ensuring privacy, security, and confidentiality of health information. These metrics can then be used by clinicians, clinical researchers, and health policy advocates to assess potential gaps in the applicability of clinical trials in real-world settings.

Our contributions

The contributions discussed in this paper are (1) formulating the problem of representativeness evaluation in RCTs as a comparison between a truly random sampling function in a target population and the actual sampling observed in the clinical trial cohort; (2) deriving new metrics for representativeness of RCT cohorts based on ML fairness metrics; (3) utilizing proposed metrics to measure subject representation of RCT cohorts with respect to a target population; (4) identifying needs, gaps, and barriers of equitable representation of various subgroups in RCT cohorts; (5) designing a tool (an R Shiny App) to automatically evaluate trial representativeness through on-demand subject stratification and distribute reports containing visualizations and explanations for different users.

METHODS AND MATERIALS

We establish a general mapping from RCT to ML fairness and then derive metrics to evaluate the population representation of RCT cohorts based on ML fairness measures.[35–41] We provide a visual representation of results with associated statistical tests to transparently communicate the quantitative results to diverse user groups.

Table 1 provides a glossary of fairness and representativeness terms used throughout the manuscript.

RCT representativeness and ML fairness

In an ML prediction model, given a feature vector x of subject from distribution $\mathcal{P}$, a binary classifier predicts if the subject is positive ($y' = 1$) or negative ($y' = 0$). The true outcome is $y \in \{0, 1\}$. We define RCT representativeness as how well the RCT cohort represents a random sampling of subjects from the specified target distribution. The target distribution can be defined based on analysis goals, for example, eligibility criteria could be considered if appropriate. In RCTs, the feature vector x is the protected attributes or subject traits; the binary classifier assigns subjects into the study cohort, where $y' = 1$ means a subject is recruited while $y' = 0$ means not recruited. y is the true random sampling result of the subject into the study from the target population.

For RCT representativeness evaluation, each individual in the target population is defined by $I = (X, y) = \big((x, x'), y\big)$, where $x \in X$ represents the protected attributes, $x' \in X$ represents the unpro-

Table 1. Glossary

Term	Definition	Example(s)
Target population	The group of people that investigators defined to be compared with the RCT cohort	US population with hypertension as defined in NHANES
Subgroup	Subset of target population that share single or multiple common baseline attribute values and thus can be distinguished from the rest	Non-Hispanic black female subjects; non-Hispanic white male subjects
Ideal rate	Proportion of subjects in a subgroup in the target population	Proportion of female subjects among those with hypertension in United States
Observed rate	Proportion of subjects in a subgroup in the RCT	Proportion of female subjects in SPRINT study
Representativeness	The similarity between an RCT sample and its target population distributions	
Protected attribute	Attributes that classify the population of a specific disease into groups that have parity in terms of health outcomes received	Age, BMI, total cholesterol
Representativeness metric	Function of disease-specific observed and ideal rates of sampling of protected subgroups to the RCT	Log disparity

Abbreviations: BMI: body mass index; NHANES: National Health and Nutrition Examination Survey; RCT: randomized clinical trial; SPRINT: Systolic Blood Pressure Intervention Trial.

tected attributes, and $y \in \{0,1\}$ is the ideal sampling of the individual by an RCT. An ideal RCT enrolls subjects i.i.d. from the target population $\mathcal{P}$. The RCT enrollment strategy can be treated as a binary classifier $\mathcal{D}(X) = y' \in \{0,1\}$, denoting the real observed decision induced by $\mathcal{D}$ on an individual i. The subgroups are defined via a family of indicator functions $\mathcal{G}$. For each $g \in \mathcal{G}$, $g(x) = 1$ means that an individual with protected attributes x is in the subgroup. For this study, we utilize protected attributes of 3 types: demographic characteristics, risk factors, and laboratory results. Here, risk factors are any study-specific covariates defined in the Table 1s of clinical trial publications relevant to the study besides demographic characteristics. The selected variables were both relevant to the study and available in the NHANES data to estimate the target distribution. Any available categorical attributes representation of the target and cohort populations could be used.

ML fairness metrics are concerned with guaranteeing similarity results across different subgroups.[42] We assume that the ideal RCT achieves statistical parity,[43] that is, subgroups are independent of outcomes ($g(x) \perp y$). Then we create metrics based on ML fairness measures of statistical parity violations. The proposed metrics also assume that the ideal sampling of a subject to the RCT and the observed sample are independent ($y \perp y'$), and the sizes and the rates of an ideal RCT and the observed trial are the same ($P(y=1) = P(y'=1)$).

The ideal and observed rates of a subgroup are $P(g(x) = s | y = 1)$ and $P(g(x) = s | y' = 1)$, respectively. The enrollment fraction of a subgroup is $P(y' = 1 | g(x) = s)$. We note by independence assumptions of ideal RCT, $P(y' = 1 | y = 1, g(x) = s) = P(y' = 1 | g(x) = s)$.

Log disparity metric for RCT

In ML fairness, the disparate impact measure is the ratio of positive rates of both protected and unprotected groups:[44]

$$\frac{P(y' = 1 | g(x) = 1)}{P(y' = 1 | g(x) = 0)}.$$

Disparate impact adopts the "80 percent rule" suggested by the US Equal Employment Opportunity Commission[45] to decide when the result is unfair:

$$\frac{P(y' = 1 | g(x) = 1)}{P(y' = 1 | g(x) = 0)} \leq \tau = 0.8.$$

The "80 percent rule" requires the selection rate of a subgroup to be at least 80% of the selection rate of the other subgroups.

As shown in the following theorem, when applied to the RCT, disparate impact reduces to an intuitive quantity based on the enrollment odds of a protected group and in the target.

Theorem 1:RCT version of Disparate Impact Metric

Based on the ideal RCT assumptions above, the disparate impact metric is equivalent to the ratio of enrollment odds of subjects of the protected group in the observed cohort to the odds of protected subjects in the ideal cohort:

$$\frac{P(y' = 1 | g(x) = 1)}{P(y' = 1 | g(x) = 0)} = \frac{\text{odds}(g(x) = 1 | y' = 1)}{\text{odds}(g(x) = 1 | y = 1)} = \frac{\text{odds}(g(x) = 1 | y' = 1)}{\text{odds}(g(x) = 1)}.$$

See Supplementary Materials for proof.

Since log odds provide advantages for ease of understanding, we propose the following metric for RCT.

Proposed Metric 1.The *Log Disparity* metric for measuring how representative of subgroup $g(x) = 1$ in observed trial y' as compared to ideal population y is

$$\log\left(\text{odds}(g(x) = 1 | y' = 1)\right) - \log(\text{odds}(g(x) = 1)).$$

In the log disparity metric, a value of 0 indicates perfect clinical equity. A value smaller than the lower threshold, $-\tau_{\text{lower}}$, implies a potential underrepresentation of a subgroup while a value greater than τ_{lower} implies a potential overrepresentation. We further add an upper threshold, τ_{upper}. A value less than $-\tau_{\text{upper}}$ implies highly underrepresentation; similarly, a value greater than τ_{upper} implies highly overrepresentation. Values between $-\tau_{\text{lower}}$ and τ_{lower} mean equitable representation.

Our metric thresholds are selected based on guidance from literature,[28,46–48] but other optimal thresholds under different criteria are allowed as inputs. We use a significance level of 0.05, a lower threshold of $-\log(0.8)$, and an upper threshold of $-\log(0.6)$.

Normalized parity metric

The ML fairness Equal Opportunity[49] metric which requires subgroups to have the same true positive rates can also be applied to RCTs.

Theorem 2:RCT version of Equal Opportunity Metric

Let ideal RCT assumptions hold and $g(x)$ be binomial random variable, then the ML fairness Equal Opportunity metric has the following equivalent form:

$$\begin{aligned} &P\left(y' = 1|g(x) = 1, y = 1\right) - P\left(y' = 1|g(x) = 0, y = 1\right) \\ &= P\left(y' = 1|g(x) = 1\right) - P\left(y' = 1|g(x) = 0\right) \\ &= \frac{P(g(x) = 1|y' = 1) - P(g(x) = 1)}{\text{var}(g(x) = 1)} P(y = 1). \end{aligned}$$

See Supplementary Materials for proof. The proportion of population in the trial, $P(y = 1)$, is extremely small and not very meaningful, thus we propose a new metric. The *Normalized Parity* metric measures the difference in rates of protected group in the trial and in the population scaled by the variance of the protected group in the target population.

Proposed Metric 2.The *Normalized Parity* metric for measuring how representative of subgroup $g(x) = 1$ in observed trial y' as compared to ideal population y

$$\frac{P(g(x) = 1|y' = 1) - P(g(x) = 1)}{\text{var}(g(x) = 1)}.$$

The proposed Log Disparity and Normalized Parity metrics have several nice properties.

1. They are easy to compute. The observed rates of each subgroup, $P(g(x) = 1|y' = 1)$, are estimated from trial data. The ideal rates and variance, $P(g(x))$ and $\text{var}(g(x))$, are estimated for the desired target population $\mathcal{P}$using surveillance datasets such as NHANES or electronic medical records (EMRs). The required estimates are robust to missing data. Individual privacy can be protected since only summary statistics are required for the proposed metrics, avoiding the pitfalls of alternative metrics requiring per subject calculations.[50]
2. Both metrics have a common interpretation for subgroups with very different background rates: 0 means that demographic parity holds, <0 means subgroup is underrepresented, and >0 means subgroup is overrepresented.
3. Statistical tests quantify the significance of observed disparities for each subgroup which take into account the RCT study size and estimation errors of the ideal assignment rate. We use a one-proportion two-tailed z-test to determine whether the observed rate is significantly deviated from the ideal population rate. We use Benjamini–Hochberg to correct for multiple comparisons across all subgroups. If the difference between observed and ideal rates is not statistically significant, the subgroup is treated as representative; otherwise, we will use metrics to quantify the subgroup representativeness. Other statistical tests could be used. See Supplementary Material for details.

Log disparity and normalized parity are both monotonically increasing functions of the observed rate for a subgroup scaled by the target rate. Log disparity offers some advantages when examining rare subgroups because it is a nonlinear function while normalized parity is a linear function, as discussed in the Supplementary Material (section: Log Disparity vs Normalized Parity). Thus, we focus on log disparity results. All Normalized Parity results are available in the supplement visualization tool.

RCT trial data

We assess the proposed methodologies on 3 real-world RCTs: Action to Control Cardiovascular Risk in Diabetes (ACCORD),[51] Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT),[52] and Systolic Blood Pressure Intervention Trial (SPRINT)[53] in BioLINCC with the ideal subgroup assignment rate calculated from individuals with matched disease conditions in NHANES. According to participants' baseline characteristics typically summarized in Table 1s of clinical trial reports, we selected 9 protected attributes. We categorize continuous variables based on the CDC (Centers for Disease Control and Prevention)-approved standards. Subject data obtained from RCTs are mapped to the existing NHANES categories. The protected attributes examined here are (1) demographic characteristics (gender, race/ethnicity, age, and education); (2) baseline risk factors [smoking status, body mass index, and systolic blood pressure (SBP)]; and (3) baseline laboratory test results [fasting glucose (FG) and total cholesterol (TC)].

The observed rates of the subgroup are calculated from the RCT data

$$P(g(x) = 1|y' = 1) = \frac{\text{numberofRCTparticipantswhosatisfied}g(x) = 1}{\text{numberofparticipantswithtargetdiseaseinRCT}}.$$

For each study, we construct all possible subgroups that can be instantiated as $g(x)$. We define 29 univariate, 109 bivariate, and 306 multivariate subgroups based on 9 protected attributes. In general, any baseline subject attributes can be selected as protected attributes in our approach.

Target population

In our experiment, we sought to evaluate how well the studies represented overall diabetic and hypertensive populations in United States as characterized by NHANES. The ideal rates from target populations ($P(g(x) = 1|y = 1)$) are calculated from NHANES 2015–2016 using the R survey() package[54] which accounts for potential bias from complex survey designs. The NHANES population selected varies based on study objectives and desired target population. To evaluate ACCORD,[51] we estimate ideal rates of subgroups of diabetic individuals in the United States using subjects who report having diabetes in NHANES, and we use subjects who report having hypertension in NHANES as the target population to evaluate ALLHAT[52] and SPRINT.[53] These criteria could be modified to consider study inclusion and exclusion criteria depending on the goals of analysis.

Since users may have better target population data that match their studies, user-provided target population datasets and multiple target files are allowed. For example, clinicians who focus on their local communities could use the community or health system population as the target to evaluate the equity of RCTs, whereas researchers who work on a global disease, the target population may be better estimated from global population datasets.

RESULTS

To demonstrate the proposed metric, we created a visualization using different colors to represent different representativeness levels in RCTs. For compact presentation, we focus on the log disparity metric. Figure 2 illustrates how the log disparity function applies to rela-

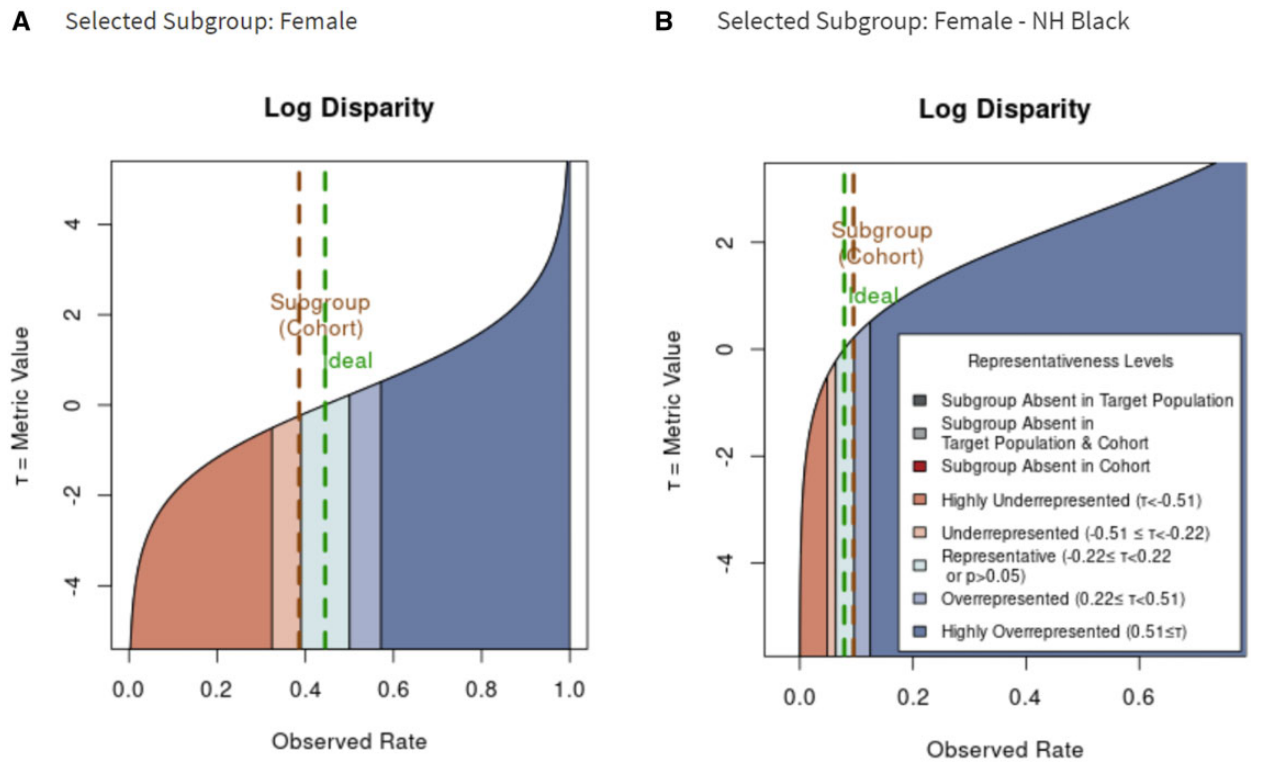

Figure 2. The shift of representativeness distribution of Log Disparity metric for different patient subgroups with type 2 diabetes in Action to Control Cardiovascular Risk in Diabetes. The green line corresponds to the ideal rate for the subgroup determined from National Health and Nutrition Examination Survey. The brown line indicates the rate actually observed. **A.** Log Disparity as function of observed rate for female subgroup. **B.** Log Disparity as function of observed rate for female non-Hispanic black subgroup.

tive common subgroups Female and Female Non-Hispanic Black in ACCORD.

As shown in Figure 2A, for women with type 2 diabetes, the ideal rate from NHANES is 0.445 while the observed RCT rate is 0.386. The observed female-subject rate falls into the light orange region, which reveals the underrepresentation of female subjects. For Figure 2B, when the subgroup of interest is changed to non-Hispanic black female participants, the ideal rate decreases to 0.079 and the observed rate becomes 0.095. Now the interested subgroup falls into the teal region, which means that non-Hispanic black female participants are equitably represented in ACCORD. This indicates the influence of protected attribute race/ethnicity on the representativeness evaluation. By comparing Figure 2A and B, we can observe that metric functions change as the ideal rate changes.

The representativeness of 29 univariate subgroups for 3 RCTs are shown in Figures 3 and 4. Dark red represents the subgroups absent from the RCT; light orange and orange indicate that subgroups are underrepresented or highly underrepresented in the RCT relative to the target population; light blue and blue specify the potentially overrepresented or highly overrepresented subgroups; teal shows the subgroup is either equitably represented or has no significant difference; dark gray indicates that no individuals with selected protected attributes exist in estimated target population; light grey indicates absent subgroup in both estimated target population and RCT.

We evaluate our ideal estimates for ACCORD, ALLHAT, and SPRINT using prior literature. For example, an estimated probability of female patients among US hypertensive population in 2015,[55] calculated through Bayes' formula, is about 47%. Comparing to the summary statistics in published literature (ie, about 47% subjects are women in ALLHAT and 36% subjects are women in SPRINT[56–59]). ALLHAT captures the gender distribution among real-world hypertensive participants while SPRINT fails to enroll enough female participants.

The color change across categories of an attribute highlights interesting trends in subject representation. Among 3 studies, only 2 attributes achieved equitable representation across all subgroups: gender in ALLHAT and TC in SPRINT. From Figures 3 and 4, we observe that current smokers, young participants, non-Hispanic Asian subjects, subjects with SBP under 130 mm Hg or FG between 5.6 and 6.9 mmol/L are frequently underrepresented. This indicates that some subgroups in the target population are missing or inadequately represented in the RCTs. The decision-making on a subject, for example, aged 40, based on the SPRINT study would require additional evidence beyond this study. Also, participants with lower education levels tend to be more underrepresented in the SPRINT while participants with higher education levels tend to be more underrepresented in the ALLHAT. This points out that potential social determinant confounders may exist in the RCT. We note, across all 3 studies, non-Hispanic black participants are overrepresented, perhaps reflecting efforts to ensure minority participation or reflecting study locations. In both hypertension RCTs, Asian subjects may have been insufficiently enrolled. This underrepresentation may also reflect study choices or locations. These trends have to be validated by analysis on more RCTs.

For subgroups defined by multiple attributes, sunburst plots better visualize the change of subgroup representation by adding additional protected attributes, as shown in Figure 5. For each type of protected attributes (ie, demographic characteristics, risk factors, and lab results), separate sunburst charts are generated since their matched population from NHANES are different.

Demographic Characteristics	Representativeness		
	ACCORD	ALLHAT	SPRINT
Gender			
Female	-0.25	-0.21	-0.68
Male	0.25	0.21	0.68
Age (Diabetes/Hypertension)			
18-44/18-39	-4.21	Absent	Absent
45-64/40-59	0.86	-0.87	-0.74
64+/59+	-0.28	1.45	1.32
Race/Ethnicity			
Non-Hispanic White	0.23	-0.73	-0.30
Non-Hispanic Black	0.32	1.04	0.96
Non-Hispanic Asian	Absent (C & TP)	-1.36	-1.63
Hispanic	-1.07	0.53	-0.16
Other/Unknown	0.14	-1.58	-1.90
Education			
Less than high school	-0.46	1.08	-2.32
High-school graduate	0.30	0.26	-0.47
Some college/Technical school	-0.06	-0.27	0.31
College degree or higher	0.14	-1.74	0.69

Figure 3. Representativeness of subgroups defined by a single protected attribute using Log Disparity for 3 real-world randomized clinical trials (RCTs). Subgroups are defined by demographic characteristics. Teal cells with a star indicate that no statistically significant difference between subgroups from the RCT and target population. Ages are in years. *Abbreviations*: C: cohort; TP: target population.

Figure 5 demonstrates log disparity results for ACCORD on demographic characteristics, ALLHAT on risk factors, and SPRINT on lab results. The interactive sunburst diagram enables users to investigate many subgroups simultaneously to identify missing or underrepresented subgroups in RCTs and NHANES. For example, young female subjects aged under 45 are missing entirely. As shown in Figure 5D, with an additional attribute FG, new subgroups such as participants with glucose $\geq$7 mmol/L are highly underrepresented for both high and normal TC. This indicates the importance of multivariable subgroup analyses in representativeness. Note that underrepresentativeness may be due to legitimate choices in the study inclusion and exclusion criteria. If desired by the user, absent subgroups in NHANES or any target populations can be estimated using smoothing techniques.

The sunburst plots explicitly address diversity, equity, and inclusion of clinical studies with respect to the target population. For instance, Figure 5B identifies the missing evidence in subgroups including any female and non-Hispanic white male subjects aged under 45. This lack of subject diversity may lead to similar results as shown for the effectiveness of Actemra on COVID-19 patients, in which the study results flipped after including more marginalized participants. Furthermore, our visualization automatically checks if the inclusion and exclusion criteria are met. Based on the criteria of SPRINT, it successfully excluded subjects with SBP under 130 mm Hg but subjects with potential impaired glucose or diabetes still existed based on the lab results.

DISCUSSION

An advantage of the proposed metrics is they provide a standardized scale for judging trial representativeness for subgroups with vastly different expected rates in the trial; for example, the estimated ideal rate of participation in the type 2 diabetes trial estimated from NHANES for subgroups of female subjects, female subjects aged over 64, Hispanic female subjects aged over 64, and Hispanic female subjects aged over 64 with high school degree are 0.445, 0.172, 0.025, and 0.006, respectively. Evaluating differences between simple rates for many subpopulations would be more challenging.

To facilitate visualizations of measured performance on clinical trials, we have incorporated a comprehensive set of fairness metrics into our prototype representativeness visualization tool using R shiny to enable researchers and clinicians to rapidly visualize and assess all potential misrepresentation in a given RCT for all possible subgroups. In our application, the number and order of the attributes for the sunburst can be changed by users; for example, instead of Figure 5B, users can visualize representativeness of subgroups for Age with further divisions by Gender and then Race/Ethnicity. With these metrics, users can rapidly determine underrepresentation of subgroups which can serve as basis for determining any limitations of the RCT. The metrics and visualizations can potentially help support evaluation of representativeness of exist-

Clinical Characteristics	Representativeness		
	ACCORD	ALLHAT	SPRINT
Cigarette-smoking status			
Current smoker	-0.91	-0.93	-1.48
Not smoke	0.91	0.93	1.48
Body-mass index group			
Underweight	-2.57	0.05★	-0.68
Normal weight	-0.26	0.22	-0.16
Overweight	0.06	0.39	0.33
Obese	0.05	-0.49	-0.20
Systolic blood pressure			
<120	-0.94	-1.95	Absent
120-129	-0.29	-1.13	Absent
130-139	0.53	-0.19	0.87
>=140	0.58	1.57	1.29
Total cholesterol			
Normal	-0.25	-1.00	0.13
High	0.25	1.00	-0.13
Fasting glucose			
<5.6	-0.56	0.93	1.32
5.6-6.9	-0.84	-1.34	-0.55
>=7	0.84	0.46	-2.10

Figure 4. Representativeness of subgroups defined by a single protected attribute using Log Disparity for 3 real-world randomized clinical trials. Subgroups are defined by clinical characteristics. Systolic blood pressure unit = mm Hg; Fasting glucose unit = mmol/L.

ing RCT cohorts, design of new RCTs, and monitoring of enrollment in ongoing RCTs. The visualization may also help healthcare providers quickly understand the applicability of RCT results to a patient in a subgroup.

Clinical trials are a key component of health equity. In the context of trial equity, underrepresentation or exclusion of disadvantaged participants may reduce opportunities to live healthy lives. Our metrics can also be applied to many types of clinical research and representativeness problems by appropriately adjusting the target population statistics based on the population of interest. Besides use with RCTs, these metrics can be easily modified to assess and visualize any disparities related to health including the distribution of medical care and different levels of living and working conditions for patients if the matching background information is available to obtain the ideal rate of each subgroup. Furthermore, our approach can be used as a frame of reference to guide the clinicians and policy-makers to make decisions with legitimate reasons and evidence. We offer user selections to dynamically control different conditions including subgroup characteristics, metric types, metric cutoffs, under which the users will make their own decisions.

The technical challenges we encountered include determining how to appropriately treat continuous variables such as age and consider inclusion and exclusion criteria when mapping RCT cohorts and NHANES-based target population. Currently, we discretize all continuous variables, with alternative approaches, such as using expected values of numerical variables and other methods applied to ML framework, left as future work. It may be desirable to further refine the target populations to adjust for missing and underrepresented subgroups due to RCT inclusion and exclusion criteria. Due to limitations in the types of information gathered in NHANES, we could not apply all eligibility criteria used in the ALLHAT, ACCORD, and SPRINT studies to define respective clinical populations for our analyses. We plan to validate our metrics by applying them to more trials and compare results with other metrics such as GIST 2.0. It can also be useful to create a method combining the proposed metrics with GIST to enable detailed subpopulation analyses of inclusion and exclusion criteria and analysis of multiple trials. Using appropriate defining target populations with eligibility criteria, these approaches can be extended to make equitable single-/multi-site enrollment planning and monitor the enrollment process to optimize

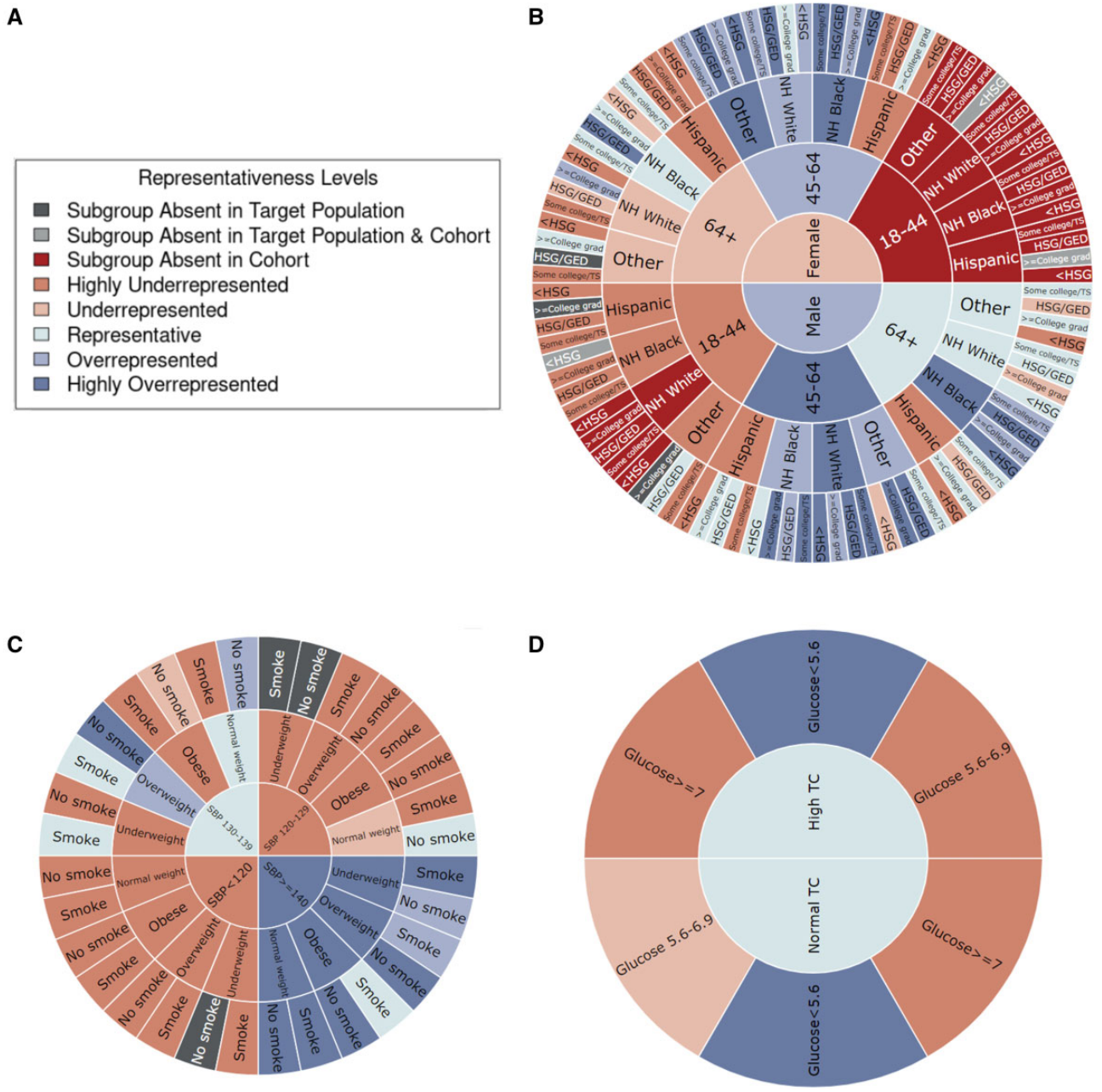

Figure 5. Representativeness results measured by Log Disparity. **A**. Color code of representativeness levels. **B**. Representativeness of Action to Control Cardiovascular Risk in Diabetes randomized clinical trial (RCT) subgroups in sunburst plot with inner to outer rings defined by demographic characteristics gender, age, race/ethnicity, and education level, respectively. **C**. Representativeness of Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial RCT subgroups in sunburst plot with inner to outer rings defined by risk factors systolic blood pressure, body mass index, and smoking status, respectively. **D**. Representativeness of Systolic Blood Pressure Intervention Trial RCT subgroups in sunburst plot with inner to outer rings defined by lab results total cholesterol and fasting glucose, respectively.

the representativeness of participants a priori and throughout the process.

CONCLUSION

Quantifying representation is important for scientific rigor and to build true equity into research designs and methods. Health equity is not just a clinical issue; it is a socioeconomic concern with broad consequences.[60–62] We developed metrics and methods to evaluate how equitably subgroups are represented in RCTs. Unlike most existing studies which focus on one protected attribute each time (eg, race) for a single disease (eg, type 2 diabetes), our proposed approach can analyze clinical trials designed for several diseases such as hypertension and type 2 diabetes, simultaneously and can additionally report representativeness of subgroups defined by multiple attributes including age and race/ethnicity. Our next steps are to utilize these metrics to monitor existing RCTs, help design new RCTs, and provide tools to disseminate findings to a variety of stakeholders

and user groups, including patients, clinicians, data scientists, and policy-makers, who will bring the discoveries into play to advance health equity.

ETHICAL APPROVAL

This manuscript was prepared using ACCORD, ALLHAT, and SPRINT Research Materials obtained from the NHLBI Biologic Specimen and Data Repository Information Coordinating Center and does not necessarily reflect the opinions or views of the ACCORD, ALLHAT, SPRINT, or the NHLBI. All methods were carried out following the NHLBI approved research plan: Equity in Clinical Trials, and all procedures were carried out in accordance with the applicable guidelines and regulations from NHLBI Research Materials Distribution Agreement. The procedures were approved by The Rensselaer IRB as IRB Review Not Required. Informed consent was obtained from all subjects by NHLBI. Data from research participants who refused to permit the sharing of their data are deleted from the repository dataset.

SUPPLEMENTARY MATERIAL

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

FUNDING

This work was primarily funded by IBM Research AI Horizons Network. All authors were supported by IBM. KPB, MQ, DMG, and OC were also supported by Rensselaer Institute for Data Exploration and Applications. KPB and OC were also supported by United Health Foundation.

AUTHOR CONTRIBUTIONS

KPB, AKD, DMG, and MAF designed and directed the project. MQ, KPB, and OC designed the model and the computational framework. MQ performed the experiments, analyzed the results, built the application, and wrote the manuscript in consultation with KPB, AKD, DMG, and MAF. All authors reviewed the manuscript.

CONFLICT OF INTEREST STATEMENT

None declared

DATA AVAILABILITY

The example ideal national patient data are calculated from the National Health and Nutrition Examination Survey (NHANES) 2015–2016 conducted by the National Center for Health Statistics (NCHS). The clinical trial data that support the findings of this study are available from Biologic Specimen and Data Repository Information Coordinating Center (BioLINCC) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are, however, available with permission of BioLINCC. The data generated during and analyzed during the current study are available in the GitHub repository, https://github.com/TheRensselaerIDEA/ClinicalTrialEquity, and via the Dryad Digital Repository at https://doi.org/10.5061/dryad.76hdr7sxf.

REFERENCES

1. Erves JC, Mayo-Gamble TL, Malin-Fair A, *et al.* Needs, priorities, and recommendations for engaging underrepresented populations in clinical research: a community perspective. *J Community Health* 2017; 42 (3): 472–80.
2. Gray TF, Cudjoe J, Murphy J, *et al.* Disparities in cancer screening practices among minority and underrepresented populations. *Semin Oncol Nurs* 2017; 33 (2): 184–98.
3. Adler NE, Rehkopf D. U.S. disparities in health: descriptions, causes, and mechanisms. *Annu Rev Public Health* 2008; 29 (1): 235–52.
4. Aristizabal P, Singer J, Cooper R, *et al.* Participation in pediatric oncology research protocols: racial/ethnic, language and age-based disparities. *Pediatr Blood Cancer* 2015; 62 (8): 1337–44.
5. Rottas M, Thadeio P, Simons R, *et al.* Demographic diversity of participants in Pfizer sponsored clinical trials in the United States. *Contemp Clin Trials* 2021; 106: 106421.
6. Krouse HJ. COVID-19 and the widening gap in health inequity. *Otolaryngol Head Neck Surg* 2020; 163 (1): 65–6.
7. Kline NS. Rethinking COVID-19 vulnerability: a call for LGBTQ+ Im/migrant health equity in the United States during and after a pandemic. *Health Equity* 2020; 4 (1): 239–42.
8. Lee S. COVID-19 amplifiers on health inequity among the older populations. *Front Public Health* 2020; 8: 609695.
9. Borno HT, Zhang S, Gomez S. COVID-19 disparities: an urgent call for race reporting and representation in clinical research. *Contemp Clin Trials Commun* 2020; 19: 100630.
10. Moore JT, Ricaldi JN, Rose CE, *et al.*; COVID-19 State, Tribal, Local, and Territorial Response Team. Disparities in incidence of COVID-19 among underrepresented racial/ethnic groups in counties identified as hotspots during June 5-18, 2020—22 states, February-June 2020. *MMWR Morb Mortal Wkly Rep* 2020; 69 (33): 1122–6.
11. Gold JAW, Rossen LM, Ahmad FB, *et al.* Race, ethnicity, and age trends in persons who died from COVID-19—United States, May-August 2020. *MMWR Morb Mortal Wkly Rep* 2020; 69 (42): 1517–21.
12. Kiene H, Hamre HJ, Kienle GS. In support of clinical case reports: a system of causality assessment. *Glob Adv Health Med* 2013; 2 (2): 64–75.
13. Zheng C, Dai R, Gale RP, *et al.* Causal inference in randomized clinical trials. *Bone Marrow Transplant* 2020; 55 (1): 4–8.
14. Roach KE. A clinician's guide to specification and sampling. *J Orthop Sports Phys Ther* 2001; 31 (12): 753–8.
15. Lim CY, In J. Randomization in clinical studies. *Korean J Anesthesiol* 2019; 72 (3): 221–32.
16. Kennedy-Martin T, Curtis S, Faries D, *et al.* A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. *Trials* 2015; 16: 495.
17. Stuart EA, Bradshaw CP, Leaf PJ. Assessing the generalizability of randomized trial results to target populations. *Prev Sci* 2015; 16 (3): 475–85.
18. He Z, Tang X, Yang X, *et al.* Clinical trial generalizability assessment in the big data era: a review. *Clin Transl Sci* 2020; 13 (4): 675–84.
19. Pradhan R, Abrahami D, Yin D, *et al.* Defining clinically relevant target populations using real-world data to guide the design of representative antidiabetic drug trials. *Clin Pharmacol Ther* 2021; 109 (5): 1219–23.
20. He Z, Chandar P, Ryan P, *et al.* Simulation-based evaluation of the generalizability index for study traits. *AMIA Annu Symp Proc AMIA Proc* 2015; 2015: 594–603.
21. Sen A, Chakrabarti S, Goldstein A, *et al.* GIST 2.0: a scalable multi-trait metric for quantifying population representativeness of individual clinical studies. *J Biomed Inform* 2016; 63: 325–36.
22. He Z, Ryan P, Hoxha J, *et al.* Multivariate analysis of the population representativeness of related clinical studies. *J Biomed Inform* 2016; 60: 66–76.
23. Li Q, Guo Y, He Z, *et al.* Using real-world data to rationalize clinical trials eligibility criteria design: a case study of Alzheimer's disease trials. *AMIA Annu Symp Proc* 2020; 2020: 717–26.
24. Murthy VH, Krumholz HM, Gross CP. Participation in cancer clinical trials race-, sex-, and age-based disparities. *JAMA* 2004; 291 (22): 2720–6.

25. ahrq.gov. Toolkit for Using the AHRQ Quality Indicators. [Internet]. 2012. https://www.ahrq.gov/patient-safety/settings/hospital/resource/qitool/index.html Accessed May 3, 2021.
26. Stuart EA, Cole SR, Bradshaw CP, *et al.* The use of propensity scores to assess the generalizability of results from randomized trials. *J R Statist Soc A* 2011; 174 (2): 369–86.
27. Tipton E. How generalizable is your experiment? An index for comparing experimental samples and populations. *J Educ Behav Statist* 2014; 39 (6): 478–501.
28. Bellamy RKE, Dey K, Hind M, *et al.* AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *IBM J Res Dev* 2018; 4: 1–15.
29. Hutchinson B, Mitchell M. 50 years of test (un)fairness: lessons for machine learning. In: ACM Conference on Fairness, Accountability, and Transparency 2019; Atlanta, GA; 2019: 49–58.
30. Srivastava M, Heidari H, Krause A. Mathematical notions vs. human perception of fairness: a descriptive approach to fairness for machine learning. In: Knowledge Discovery and Data Mining 2019; Anchorage, AK; 2019: 2459–68.
31. National Center for Health Statistics. National Health and Nutrition Examination Survey Data. [Internet]. 2016. https://wwwn.cdc.gov/nchs/nhanes/ Accessed May 3, 2021.
32. Bennett JA. The consolidated standards of reporting trials (CONSORT): guidelines for reporting randomized trials. *Nurs Res* 2005; 54 (2): 128–32.
33. Falci SGM, Marques LS. CONSORT: when and how to use it. *Dental Press J Orthod* 2015; 20 (3): 13–5.
34. Welch VA, Norheim OF, Jull J, *et al.*; CONSORT-Equity and Boston Equity Symposium. Research methods reporting consort-equity 2017 extension and elaboration for better reporting of health equity in randomised trials. *BMJ* 2017; 359: j5085.
35. Saha D, Schumann C, McElfresh DC, *et al.* Measuring non-expert comprehension of machine learning fairness metrics. In: International Conference on Machine Learning 2020; Vienna, Austria; 2020; 119: 8377–87.
36. Cynthia D, Moritz H, Toniann P, *et al.* Fairness through awareness. In: Innovations in Theoretical Computer Science Conference 2012; Cambridge, MA; 2012: 214–26.
37. Kearns M, Neel S, Roth A, *et al.* Preventing fairness gerrymandering: auditing and learning for subgroup fairness. In: International Conference on Machine Learning 2018; Stockholm, Sweden; 2018; 80: 2564–72.
38. Agarwal A, Beygelzimer A, Dudik A, *et al.* A reductions approach to fair classification. In: International Conference on Machine Learning 2018; Stockholm, Sweden; 2018; 80: 60–9.
39. Saxena NA, Huang K, DeFilippis E, *et al.* How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In: AAAI/ACM Conference on AI, Ethics, and Society 2019; Honolulu, HI; 2019: 99–106.
40. Ustun B, Liu Y, Parkes D. Fairness without harm: decoupled classifiers with preference guarantees. In: International Conference on Machine Learning 2019; Long Beach, CA; 2019; 97: 6373–82.
41. Vasileva MI. The dark side of machine learning algorithms: how and why they can leverage bias, and what can be done to pursue algorithmic fairness. In: Knowledge Discovery and Data Mining 2020; San Diego, CA; 2020: 3586–7.
42. fairmlbook.org. Fairness and Machine Learning. [Internet]. 2019. https://fairmlbook.org/ Accessed May 3, 2021.
43. Du M, Yang F, Zou N, *et al.* Fairness in deep learning: a computational perspective. *IEEE Intell Syst* 2020; 1–1. doi: 10.1109/MIS.2020.3000681.
44. Feldman M, Friedler SA, Moeller J, *et al.* Certifying and removing disparate impact. In: Knowledge Discovery and Data Mining 2015; Sydney, Australia; 2015: 259–68.
45. Roth PL, Bobko P, Switzer F. Modeling the behavior of the 4/5ths rule for determining adverse impact: reasons for caution. *J Appl Psychol* 2006; 91 (3): 507–22.
46. Beutel A, Chen J, Doshi T, *et al.* Putting fairness principles into practice: challenges, metrics, and improvements. In: AAAI/ACM Conference on AI, Ethics, and Society 2019; Honolulu, HI; 2019: 453–9.
47. Fish B, Kun J, Lelkes ÁD. A confidence-based approach for balancing fairness and accuracy. In: SIAM International Conference on Data Mining 2016; Miami, FL; 2016: 144–52.
48. Radovanović S, Petrović A, Delibašić B, *et al.* Making hospital readmission classifier fair—what is the cost? In: Central European Conference on Information and Intelligent Systems 2019; Varaždin, Croatia; 2019: 325–31.
49. Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. In: Conference on Neural Information Processing Systems 2016; Barcelona, Spain; 2016: 3323–31.
50. Rocher L, Hendrickx JM, de Montjoye YA. Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 2019; 10 (1): 3069.
51. The ACCORD Study Group. Action to control cardiovascular risk in diabetes (accord) trial: design and methods. *Am J Cardiol* 2007; 99: 21i–33i.
52. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the antihypertensive and lipid-lowering treatment to prevent heart attack trial (ALLHAT). *JAMA* 2002; 288: 2981–97.
53. The SPRINT Research Group. A randomized trial of intensive versus standard blood-pressure control. *N Engl J Med* 2015; 373: 2103–16.
54. Lumley T. Analysis of complex survey samples. *J Stat Soft* 2004; 9 (8): 1–19.
55. Fryar CD, Ostchega Y, Hales CM, *et al.* Hypertension Prevalence and Control Among Adults: United States, 2015–2016. [Internet]. 2017. https://wwwn.cdc.gov/nchs/nhanes/ Accessed May 3, 2021.
56. Oparil S, Davis BR, Cushman WC, *et al.*; ALLHAT Collaborative Research Group. Mortality and morbidity during and after ALLHAT: results by gender. *Hypertension* 2013; 61 (5): 977–86.
57. Einhorn PT, Davis BR, Massie BM, *et al.*; ALLHAT Collaborative Research Group. The Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) heart failure validation study: diagnosis and prognosis. *Am Heart J* 2007; 153 (1): 42–53.
58. Delles C, Currie G. Sex differences in hypertension and other cardiovascular diseases. *J Hypertens* 2018; 36 (4): 768–70.
59. Cohen DL, Townsend RR. Which patients does the SPRINT study not apply to and what are the appropriate blood pressure goals in these populations? *J Clin Hypertens (Greenwich)* 2016; 18 (5): 477–8.
60. Engelgau MM, Zhang P, Jan S, *et al.* Economic dimensions of health inequities: the role of implementation research. *Ethn Dis* 2019; 29 (Suppl 1): 103–12.
61. Esposito M, Larimore S, Lee H. Aggressive Policing, Health, and Health Equity. [Internet]. 2021. https://www.healthaffairs.org/do/10.1377/hpb20210412.997570/full/ Accessed May 3, 2021
62. Östlin P, Schrecker T, Sadana R, *et al.* Priorities for research on equity and health: towards an equity-focused health research agenda. *PLoS Med* 2011; 8 (11): e1001115.