Original Research

Utilization of geospatial distribution in the measurement of study cohort representativeness

Keith Feldman [a,b,*], Natalie J. Kane [c], Stacey Daniels-Young [d], Bruce Reed [d], Jessica Welch [d], Lauren Fitzpatrick [e], Mark A. Hoffman [c], Andrea Bradley-Ewing [a], Elin Grundberg [b,f]

[a] Health Services and Outcomes Research, Children's Mercy Kansas City, 2401 Gillham Road, Kansas City, MO 64108, USA
[b] Department of Pediatrics, University of Missouri-Kansas City School of Medicine, 2411 Holmes Street, Kansas City, MO 64108, USA
[c] Research Informatics, Children's Mercy Kansas City, 2401 Gillham Road, Kansas City, MO 64108, USA
[d] Children's Mercy Research Institute Community Advisory Board, Children's Mercy Kansas City, 2401 Gillham Road, Kansas City, MO 64108, USA
[e] Center for Children's Healthy Lifestyles & Nutrition, Children's Mercy Kansas City, 2401 Gillham Road, Kansas City, MO 64108, USA
[f] Genomic Medicine Center, Children's Mercy Kansas City, 2401 Gillham Road, Kansas City, MO 64108, USA

ARTICLE INFO

ABSTRACT

Objective: The ability to apply results from a study to a broader population remains a primary objective in translational science. Distinct from intrinsic elements of scientific rigor, the extrinsic concept of generalization requires there be alignment between a study cohort and population in which results are expected to be applied. Widespread efforts have been made to quantify representativeness of study cohorts. These techniques, however, often consider the study and target cohorts as monolithic collections that can be directly compared. Overlooking known impacts to health from socio-demographic and environmental factors tied to individual's geographical location, and potentially obfuscating misalignment in underrepresented population subgroups. This manuscript introduces several measures to account for geographic information in the assessment of cohort representation.
Methods: Metrics were defined across two themes: First, *measures of recruitment,* to assess if a study cohort is drawn at an expected rate and in an expected geographical pattern with respect to individuals in a reference cohort. Second, *measures of individual characteristics,* to assess if the individuals in the study cohort accurately reflect the sociodemographic, clinical, and geographic diversity observed across a reference cohort while accounting for the geospatial proximity of individuals.
Results: As an empirical demonstration, methods are applied to an active clinical study examining asthma in Black/African American patients at a US Midwestern pediatric hospital. Results illustrate how areas of over- and under-recruitment can be identified and contextualized in light of study recruitment patterns at an individual-level, highlighting the ability to identify a subset of features for which the study cohort closely resembled the broader population. In addition they provide an opportunity to dive deeper into misalignments, to identify study cohort members that are in some way distinct from the communities for which they are expected to represent.
Conclusion: Together, these metrics provide a comprehensive spatial assessment of a study cohort with respect to a broader target population. Such an approach offers researchers a toolset by which to target expected generalization of results derived from a given study.

1. Introduction

The Department of Health and Human Service defines research as a systematic investigation "designed to develop or contribute to *generalizable* knowledge" [1]. Distinct from the intrinsic considerations of scientific rigor and reproducibility, often considered foundational elements of research, generalization (also referred to as external validity) takes a broader, extrinsic viewpoint, centering on the capacity of study findings to remain applicable in new contexts. For studies involving human subjects, this often manifests in the ability of results and conclusions drawn from a given sample (cohort) to apply across a broader population from which it was drawn [2].

Despite a multitude of published conceptual models and frameworks establishing best practices for research [3–5], successful generalization

remains a widespread challenge [6]. Historically, generalization issues have been attributed to sampling bias and unmeasured confounding induced by non-randomized and observational studies. However, it is now understood that poor external validity is common, even in gold-standard randomized control trials [7]. Researchers have since explored factors underlying this phenomenon, identifying several contributing factors including differences across study procedures (e.g., measurement), logistics (e.g., timing), and personnel [8]. Yet, it is consistently recognized that failure to generalize is often attributed to a misalignment between the study cohort and population in which the results are expected to apply.

From a statistical perspective, this misalignment can be viewed as a consequence of a non-representative sample, one in which a study cohort fundamentally differs from the broader population [9]. To address this, researchers have proposed a diverse set of methodological approaches. Some, focusing on generalizability of statistical relationships, have introduced analytical techniques to improve the reliability of inference in both the randomized and non-randomized context [10–12]. Others have posited the lack of alignment (and subsequent failure to generalize) is driven by poor representation of various populations in the study cohort as compared to the target population, focusing on measurement of subgroup heterogeneity and development of *post-hoc* corrections [13,14]. Together these efforts have sparked an emerging field related to fairness and bias of study results. Particularly in predictive modeling, such imbalances have been shown to significantly impair model performance and must be addressed through a similar set of specialized fitting procedures and post-hoc adjustments [15,16].

Together these efforts have pushed the field forward, as evidenced by a prominent systematic review of over 150 studies that utilized generalizability measures to assess clinical trial populations between 1985–2019 [17]. Efficacy of these approaches, however, is tied not only to their technical execution, but also to the comparison of the study cohort with a meaningful set of data representing the theoretic generalized population. This comparison is a non-trivial technical task, one made more complex by a natural degree of variability expected across individuals' socio-demographic, clinical, and social/structural determinants of characteristics. Recognizing that the inclusion criteria of a study can only define the population *expected* to be enrolled in a study, the majority of published approaches have leveraged large databases to generate reference cohorts (e.g., eligible but non-recruited individuals) from which to directly assess deviations to the study cohort results. This approach has been widely successful, with propensity-based models remaining the state-of-the-art in identifying and adjusting for multifactorial cohort imbalances (i.e. all older individuals were male).

Unfortunately, by considering the study and reference cohorts as monolithic pools of data that can be directly compared, these techniques overlook a critical element known to impact individual health: *geographical location*. To date, the developed approaches have largely ignored bias introduced as a result of non-uniform sampling distributions that include individuals with highly distinct social determinants. In turn, this can obfuscate misalignment between underrepresented groups in the study and target populations, and precluding correction with the previously highlighted techniques.

The techniques presented in this manuscript are designed to address exactly this. First, we outline a series of metrics to quantify the geographic representativeness of recruitment for a study cohort with respect to the distribution of individuals from a reference cohort. We then introduce a series of measures to quantify representativeness of a study cohort's individual-level characteristics that account for geospatial proximity. As a case study, these measures are applied to an ongoing clinical study examining asthma in Black/African American patients treated at a large pediatric hospital in the Midwestern United States. Based on these results, we provide a discussion around the interpretation of metrics values, as well as considerations for their use in study recruitment and analysis. We conclude by outlining limitations of

geospatial techniques and discussion of future work.

Statement of Significance	
Problem	Findings from research studies often do not generalize to individuals outside the study cohort.
What is Already Known	Successful generalization requires a study cohort be representative of the population to which results are expected to be applied. Existing techniques to quantify representativeness often compare the study and target cohorts as monolithic pools of data. Overlooking socio-demographic and environmental factors tied to geographical location.
What This Paper Adds	This manuscript expands on existing measures of cohort representativeness. Introducing a series of metrics to account for geospatial distribution, both with respect to macro-level patterns of recruitment, and measurement of individual-level characteristics.

2. Methodology

To account for geospatial distribution in the measurement of cohort representativeness, this work defined metrics across two themes: First, Measures of Recruitment, addressing the question *are individuals in a study cohort drawn at an expected rate and in an expected geographical pattern with respect to those in a reference cohort.* Second, Measures of Individual Characteristics, addressing the question *does the study cohort accurately reflect the sociodemographic, clinical, and geographic diversity observed across a reference cohort.* Details of the metrics comprising each theme are described below.

2.1. Co-development of metrics

As part of metric development, we recognized a risk of perpetuating biases by characterizing individuals based on geographic factors. In an effort to undertake this work in a responsible manner, this project recruited three members of the Children's Mercy Research Institute Community Advisory Board to become part of the study team. These individuals brought diverse experiences as parents, caregivers, and community leaders, together providing consultation on more than 30 research projects. For this work, they provided critical insight on the collection and utilization of specific data, development of metrics, and the presentation of results. Throughout the following sections, specific references to this interaction are highlighted.

2.2. Measures of recruitment

Recognizing the importance of improving access and representation in research, studies now often recruit across large geographic areas. As such, the comparison of sociodemographic characteristics alone can mask potential bias, as similar sociodemographic profiles may exist in multiple distinct geographic areas. Measuring and assuring representation of subjects across each of these areas is itself a critical factor in producing generalizable findings. The metrics proposed here are designed to provide a comprehensive spatial assessment of recruitment between the set of individuals from a study (S) and reference (R) cohort. We also define person (p) and region (G_i) as the set of individuals who reside within a geographic region *i*.

2.2.1. Coverage and recruitment rate

Recruitment dispersion and balance can be most directly measured as a percent overlap between the unique geographic areas from which study cohort is drawn and those in which individuals in the reference cohort exist (Eq. (1)). This measure, deemed *Coverage,* can be calculated at any geospatial granularity (Counties, Census Block groups, etc.) as defined by a user.

$$\mathbb{S} = \{G_i | \exists p[p \in S \wedge p \in G_i]\}$$
$$\mathbb{R} = \{G_i | \exists p[p \in R \wedge p \in G_i]\} \tag{1}$$
$$Coverage = \frac{|\mathbb{S}|}{|\mathbb{R}|}$$

Eq. (1): Percentage of unique regions for individuals in study cohort (S) and reference cohort (R). $\mathbb{S}$ and $\mathbb{R}$ represent the set of geographic regions with individuals in S and R, respectively.

Expanding on coverage, we next introduce a measure of *Recruitment Rate ($\mathbb{P}_i$),* defined as the ratio between the number of individuals recruited from geography (G_i) and the number of reference individuals in that geography (Eq. (2)). This value can be aggregated across the set of all geographic areas (n) for which reference individuals exist to summarize recruitment averages and variability.

$$\mathbb{P}_i = \frac{|\{p | p \in S \wedge p \in G_i\}|}{|\{p | p \in R \wedge p \in G_i\}|}$$
$$MeanRecruitment = \frac{1}{n}\sum_i \mathbb{P}_i \tag{2}$$

Eq. (2): Recruitment rate for a geographic region (i), as number of individuals from study cohort (S) in i, normalized by number of individuals from reference cohort (R) in i. Mean recruitment is an average recruitment across all areas (i).

2.2.2. Spread and hot spots

Next, to better identify situations in which high-coverage recruitment represents a limited geographic area, we calculate a measure of *Spread,* defined as the total geographic land area of regions that comprise members of the study cohort (Eq. (3)). Such a measure can be further contextualized by normalizing the area of a region by its total population.

$$Spread = \sum_{\forall G \in \mathbb{S}} Area(G) \tag{3}$$

Eq. (3): Total land area (Area) for the set of geographic regions covered by individuals in study cohort ($\mathbb{S}$).

Rounding out the measure of spread, we add a formal measure of geographic clustering with a hot spot analysis. Similar to how recruitment rate can be used to identify imbalances in the coverage, spatial hot spot analysis can be used to identify areas that have higher concentration of recruited individuals compared to the expected number given a random distribution. For this work, we utilized the Getis-Ord Gi* algorithm [18] provided by ArcGIS Pro v.3.1.2 (Redlands, CA) [19].

2.2.3. Sensitivity analysis

It is important to recognize that given finite sample sizes, geographic areas with small reference cohorts may introduce variability into the proposed metrics (e.g., census tracts with only 1 eligible individual, contribute either a large positive or negative effect based on if that individual was recruited – 100% vs. 0%, respectively). To improve the stability, sensitivity analyses should be performed to identify the minimal size of the reference cohort in a given geography for inclusion.

2.2.4. Expected values of recruitment

As the study cohort represents a subset of the broader reference cohort, we would not expect recruitment metrics (coverage, recruitment or spread) from *any* sample to equal that of the corresponding reference cohort. This is particularly true for smaller studies, in which cohorts capture only a fraction of the reference area due to limited sample sizes. While the absolute values of these metrics provide an overview of recruitment balance, which is critical for understanding potential biases stemming from overall geographic variance, we can also contextualize the magnitude of recruitment metrics. Specifically, we can address the question, *what would be the expected value of a given metric for a random subset of the reference cohort for a study of size (n).* To do so, we introduce

normalized measures of *expected* representativeness. For each of the above metrics, 10,000 random samples are drawn from the reference cohort equal in size to the study cohort and the metric computed treating this as the study cohort. These samples are then used to calculate the expected distribution of the metric, on which we can compare the observed value from the true study cohort.

2.3. Measures of individual characteristics

We next introduce a set of metrics designed to measure alignment between individual-level data for those who are part of the study cohort and those in the broader reference cohort while also considering their geographic proximity. For these metrics, we employ both intra-individual measures of distance, as well as aggregate measures of alignment leveraged from case-control studies. A conceptual overview of both methods can be found below, while more detailed pseudocode can be found in Appendix A.

2.3.1. Distance-based comparisons

At a granular level, alignment between study and reference cohorts can be measured as a similarity between the characteristics of individuals comprising each group. To calculate this, we compute the pairwise distance between each study cohort member and all reference individuals from their geographic region (again assigned at any standardized granularity). Distances are calculated against available individual-level features in both cohorts (age, sex, acuity, etc.) using real-valued metrics (e.g., cosine) after hot-encoding discrete data, or by utilizing mixed-type measures such as Gowers distance [20] on heterogenous datasets directly.

For a given study cohort member, the average of distance to the associated reference subjects is then contextualized using a series of bootstrap samples drawn from the reference cohort outside the individual's geographic area (equivalent in size to the number of true reference subjects in the region of the study cohort member). The average distance between the study individual and this resampled reference group can be computed. The process is repeated for many iterations to estimate an expected average distance of an individual to a random cohort. The distance to the true refence cohort can then be compared to this sampling distribution using a standard z-score.

2.3.2. Case-control approach

While the distance-based measures can be averaged across all individuals in the study cohort to obtain a summary measure, more easily interpreted aggregate measures of deviation between the study and reference cohorts can be obtained by utilizing established techniques developed for comparing the "balance" (i.e. alignment) of cohorts in case-control studies.

The design of this metric is as follows: For each study cohort member, 1 (or more) random individuals are drawn from the reference cohort in the same geographic area. These pairs are treated as "matches" and can directly utilize established case-control balance metrics. For this work we focus on standardized mean difference, and maximal proportion difference for continuous and nominal data respectively [21–23]. By repeating the sampling in a bootstrap approach, it is possible to calculate an estimation of overall alignment of a given study cohort to the broader reference cohort.

2.3.3. Sensitivity analysis

As in the recruitment metrics, stability of both individual-level measures can be tuned by assessing a minimum number of reference cohort individuals within a given geographic area to assure sufficient cohort sizes for distributional estimates. For these metrics, sensitivity analyses can also consider the identification and trimming of highly distinct reference individuals, by comparing individuals in the reference cohort to others from the matching geography. Exclusions can be done directly in the distance-based comparison, or through the use of calipers

in the case-control approach.

2.3.4. Global comparisons

Although the aforementioned metrics can directly address the question of how representative study cohort members are of others in the geographic areas in which they are drawn, we note a potential limitation. Comparing only individuals within the same geographic regions may draw an arbitrary delineation in space. Two subjects only a small absolute distance apart may be separated by a geographic boundary and thus not considered in alignment measures despite providing potentially relevant information. To address such challenges, we introduce extensions of the previously described methods to leverage the entire reference cohort. For the *distance-based comparison*, the pairwise distance between each study cohort member is performed against all individuals in the reference cohort. This distance is inversely weighted proportional to either (A) geographic distance between the two entities, or (B) by a proxy measure of community-based sociodemographic similarity (e.g., County Health Rankings [24]). For the *case-control comparisons,* we continue to utilize random sampling, but treat the inverse distance as a sampling weight to further adjust the traditional balance estimates.

Expanding the calculations to consider individuals across multiple geographic regions has two additional benefits. First, it allows for improved estimates of representations for individuals in the study cohort from regions with limited reference cohorts. Second, it allows for the inclusion of data captured at a population level (e.g., median household income). By focusing within a geography, data at this level would show no variability, yet across all individuals, these data can be considered in addition to individual data.

3. Case study

To demonstrate how the proposed methodology can be used to quantify geospatial representativeness, each of the aforementioned methods were applied to an ongoing NIH funded project; the *Social Epigenomics and Asthma* (SEA) study.

All data and protocols for the SEA study and this secondary analysis were approved by Children's Mercy Kansas City (CMKC) IRB (protocol #00001292) with a waiver of consent. Linkage of retrospective CMKC records to identify the reference cohort was covered under repository protocol (#00001981).

3.1. Study and reference cohorts

Study Cohort: The SEA study enrolled English-speaking families self-identified as Black/African American, who presented with a child (aged 18 and under) to CMKC Adele Hall emergency department or inpatient setting with acute respiratory symptoms related to an asthma diagnosis. Enrolled families completed extensive surveys (e.g., chronic stress, asthma control) along with nasal and buccal swabs for genomics assessments. This work focuses on subjects recruited between 03/15/2021 and 11/08/2022.

Reference Cohort: Defined as patients eligible, but not recruited for the SEA study. The cohort was derived from the CMKC electronic health record (EHR), identifying patients with (A) historical asthma diagnosis (B) acute care visit during the SEA recruiting period and (C) demographic records identifying race as Black/African American. Additionally, the CMKC health system is a major referral center, drawing patients from all 50 US States, and from 14 countries. For this example, we focused on those individuals from the regional Total Service Area (TSA, ~70% of all patients), which includes Jackson, Clay, and Platte Counties in Missouri, and Johnson and Wyandotte Counties in Kansas [25].

3.2. Data elements

The proposed metrics and techniques are designed to be generalizable for any data shared between the reference and study cohort at either an individual- or population-level, allowing study teams to best define those relevant to their cohort. As part of this manuscript, we demonstrate their use with an array of logistical, demographic, and clinical variables extracted from the EHR and several public data sources.

For all patients, both those in the study cohort and those in the reference cohort, these include:

Geospatial Location: Residential addresses for both study participants and reference cohort were extracted from the EHR and geocoded using the StreetMaps Premium package in ArcGIS Pro. For reliability, only those addressees coded with a geocoding score >90 were retained. Additionally, suggested by our community-based team members, an estimate of total distance between the individual's address and recruitment location (CMKC Adele Hall) was also computed via a HIPAA compliant local implementation of the DeGAUSS GeoMarker tool [26].

Demographics: For the first eligible encounter within the study period (reference cohort) or recruitment visit (SEA cohort) we estimated age at the time of encounter (encounter date v. date of birth), and primary insurance. We also captured sex, as well as self-reported race and ethnicity. Race and ethnicity were harmonized to the OMB data standards and utilized the approved sixth category to include "*other*" when specified by the individual. Individuals with multiple selected races were categorized as "multiracial" to preserve independence of categories. Insurance was categorized into *Commercial, Medicaid, Self-Pay, Combination (Commercial + Medicaid), Other Insurance, and Unknown.*

Clinical Data: All historical clinical diagnoses (rather than billing) were extracted from the EHR. Asthma diagnosis codes (*defined as ICD10 J45.*, ICD9 493.*,* and/or *internal diagnosis code description that included the term Asthma*) were used to identify first date of diagnosis (to estimate history of asthma), as well as identify acute visits during the study period with respiratory symptoms related to an asthma diagnosis. Additionally, these data were used to compute a validated pediatric comorbidity index [27] for all visits during the study period. Through discussions with community team members, we added an additional metric to represent the count of eligible visits during the study period, as a proxy for burden. It was felt that families for whom the first encounter was their first acute visit compared to those with recurrent visits may differ in a measurable way.

Population-Level Indicator Data: Census tract identifiers were collected from the U.S. Census Bureau API using the R TidyCensus package [28] and linked using a spatial join based on the geocoded location (latitude and longitude) of residential street address. Decennial Census geographic identifiers were used to join patients with the census tract of the Child Opportunity Index (COI), utilized as a sociodemographic indicator for weighting in the individual-level *global* comparisons. Note, location was mapped to the 2010 census tracts, as the COI data is only available for that version of the census tract geography. Additionally, as part of our DeGAUSS geo-mapping, social and economic data linked to the census tracts were extracted including: fraction of population 25 and older with educational attainment of at least high school graduation, median household income in the past 12 months, and fraction of vacant houses. Data were linked to most recent DeGAUSS API and represent information from the 2018 ACS data.

All data collection and processing were done using R v.4.3.1 and RStudio v.2023.09.1 [29]. In addition to geocoder tools previously described, some analytical methods (*Measures of Individual Characteristics)* and cohort processing utilized Python 3.8 and SciPy v1.7.3 [30].

Code availability: Underlying code for this study is not publicly available but will be made available to qualified researchers on reasonable request to the corresponding author.

4. Results

In total 400 participants were recruited into the SEA study during the specified period. During this timeframe, CMKC had encounters with 394,189 additional unique patients across all specialties and clinics. Of those, 42,702 were identified as having a history of asthma, with 12,699 also providing a self-reported race of "Black or African American". From this cohort, 2757 patients were identified as having at least 1 encounter for an asthma exacerbation during the study period, representing a complete reference cohort (eligible but not recruited for SEA study). Finally, filtering both cohorts to only those with reliable geocoded address in TSA, we reach a final study cohort of 380 and a reference cohort of 2553; representing 95% and 93% of possible patients, respectively. A descriptive summary of the data elements for the study and reference cohorts can be found in Table 1.

4.1. Measures of recruitment

Looking first to the overall geospatial distribution of the study and reference cohorts, Fig. 1 illustrates the frequency of individuals in each cohort broken out by the set of census tracts comprising the CMKC TSA.

Table 1
Overview of the study and reference cohorts. All continuous data reported as mean (SD), nominal data as n = (%).

		Cohort	
		Study	Reference
	n	380	2553
Demographics	**Age (Years)**	6.21 (4.35)	8.06 (5.18)
	Sex		
	Female	173 (45.53)	1095 (42.89)
	Male	207 (54.47)	1458 (57.11)
	Insurance Type		
	Commercial	45 (11.84)	498 (19.51)
	Commercial & Medicaid	23 (6.05)	163 (6.38)
	Medicaid	303 (79.74)	1800 (70.51)
	Self-Pay	8 (2.11)	64 (2.51)
	Unknown	1 (0.26)	28 (1.10)
	Ethnicity		
	Hispanic/Latino	14 (3.68)	30 (1.18)
	Non-Hispanic/Non-Latino	364 (95.79)	2523 (98.82)
	Unknown	2 (0.53)	
Individual-Level Data	**Acute Care Visits in Study Period**	4.22 (4.19)	3.12 (2.37)
	Pediatric Comorbidity Index	0.52 (0.48)	0.80 (0.68)
	Distance to CMKC (km)	9.99 (6.80)	13.46 (7.85)
Population-Level Data (within Census Tract)	**COI National (z-score)**	−0.04 (0.03)	−0.02 (0.04)
	Census Population Estimate	3262.42 (1513.10)	3745.60 (1667.54)
	Median Income	38527.20 (15046.22)	47664.30 (22505.30)
	Percent High School Education	82.87 (9.30)	86.48 (9.34)
	Percent Below Poverty Line	25.51 (12.54)	20.45 (13.10)
	Percent of Vacant Houses	18.18 (9.76)	14.00 (9.33)

4.1.1. Coverage and recruitment rate

In total, study participants were drawn from 149 unique census tracts, while the reference cohort covered 380. All tracts with study participants also included members of the reference cohort, resulting in an *absolute coverage* of approximately 39.1%. The expected number of unique census tracts for a study cohort of this size was ~190, providing an expected coverage of 78.5%.

To better understand how this coverage was allocated across the region, Fig. 2 presents the recruitment rate for each census tract. For stability and as an example of the methodology, an absolute recruitment rate was calculated for only those regions in which there were more than 3 individuals in the reference cohort (n = 202), resulting in a mean rate of 10.9% (SD 11.9 %). The expected recruitment rate for the same cohort was found to be similar at 13.0%. More nuanced assessments of recruitment can be performed by isolating those census tracts from which a study cohort subject was already recruited, with a mean (absolute) rate of 18.4% (SD 10.1%).

4.1.2. Spread and hot spots

The 149 unique tracts represent a relatively small geographic absolute area of just 234.47 sq. miles. This represents only about 23 % of the total coverage of the area covered by the total reference cohort (1000.27 sq. miles). From this population, we note a cohort of this size produces an expected spread of 420.30 sq. miles (SD: 48.60).

Further, to test for statistical clustering in the study recruitment (patterns illustrated in the study cohort panel of Fig. 1A), the *Getis-Ord Gi* hot spot analysis* was performed, utilizing an inverse distance spatial relationship and Euclidean distance. Several areas of clustered recruitment were identified. Of note, all such patterns were associated with regions of high recruitment. No specific pattern was identified for regions with low recruitment (cold spots). These results are overlaid to the TSA in Fig. 3.

4.2. Measures of individual characteristics

At the individual level, we highlight the complementary nature of the case-control and distance-base methodologies in robustly describing representativeness of the study cohort. In this sample experiment, we assess demographic and individual-level characteristics (age, sex, ethnicity, insurance, comorbidity score, distance to the hospital, and eligible visit count). Again, we focus on the subset of patients for which there were at least 3 reference patients in a matching geographic area.

First, through the series of 500-bootstrapped matching iterations, Table 2A presents an estimated alignment (balance) between each of the characteristics of the study and reference cohorts. For completeness, we illustrate how representativeness can be calculated for the study cohort utilizing the global weighting approach. Weights were set as the difference between COI of the study cohort member and that of each member of the reference cohort, and we expand the evaluated characteristics to include population-level indicators (fraction of population with high school education, median home income, fraction of vacant houses). Alignment scores for each variable can be found in Table 2B. Note: while this approach can be used on all individuals, we keep the cohort consistent to that in the within-geography analysis. Results for those excluded from within-geography analyses (reference cohort size ≤ 3) in Appendix B and were found to be generally in line with individuals in Table 2B.

While these values provide an aggregated measure of alignment across the study cohort, we then use the distance-based approach to identify specific outlier members of the study cohort. Second, we compute a z-score between the calculated distance (*Gowers*) of each study cohort member to the reference cohort subgroup in the matching geography, as compared to distribution of 500 mean-distances to a random sample of reference cohort. Using these z-scores, we isolate outlier study cohort members. In this experiment, we identify those in which z-scores that indicate the study cohort member was statistically
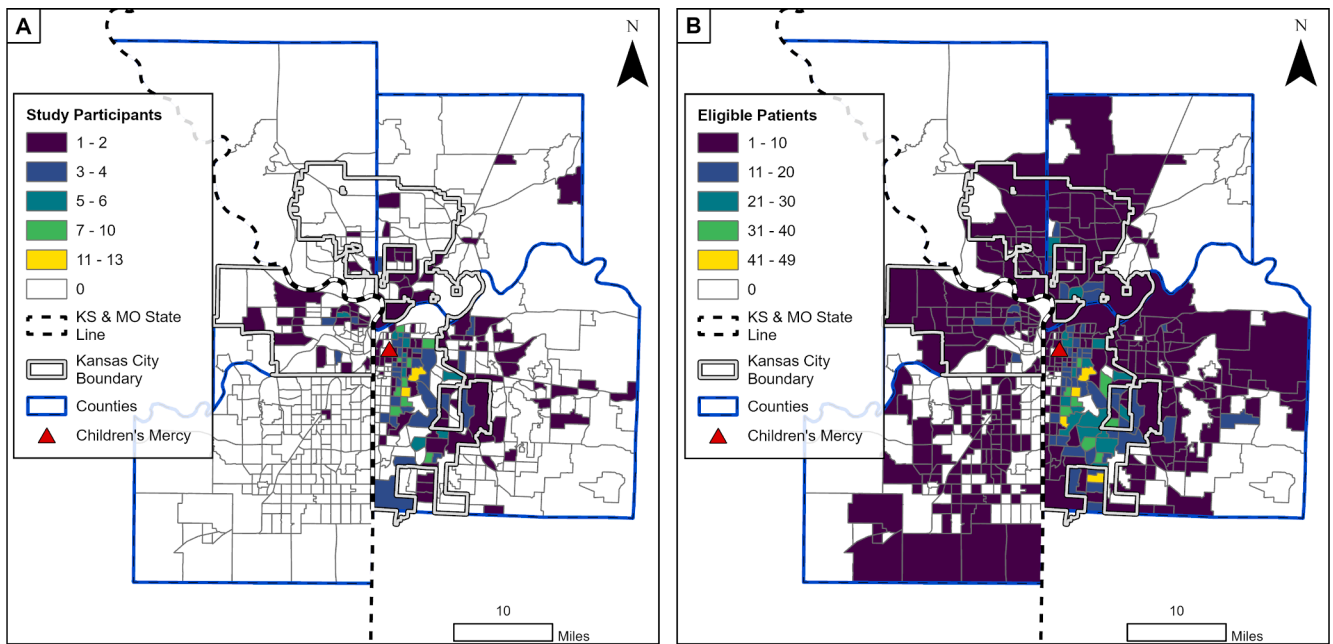
Fig. 1. Recruitment Distribution: Count of patients from the SEA study cohort (Panel A) and reference cohort (Panel B) within each of the 2020 US Census Tracts for the catchment area of Children's Mercy Kansas City.
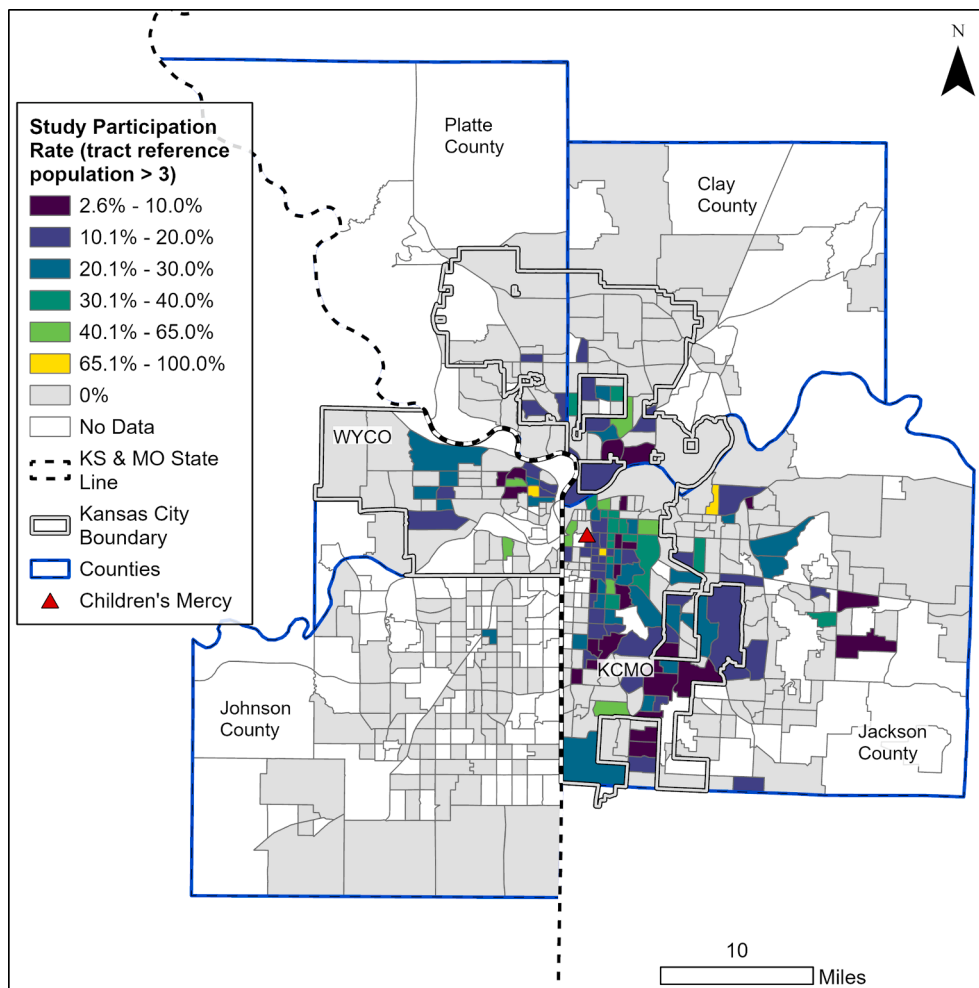
Fig. 2. Recruitment Rate: calculated ratio of SEA patients to eligible patients within a given census tract. For stability, considered census tracts include only those with at least 3 eligible patients from the reference cohort.
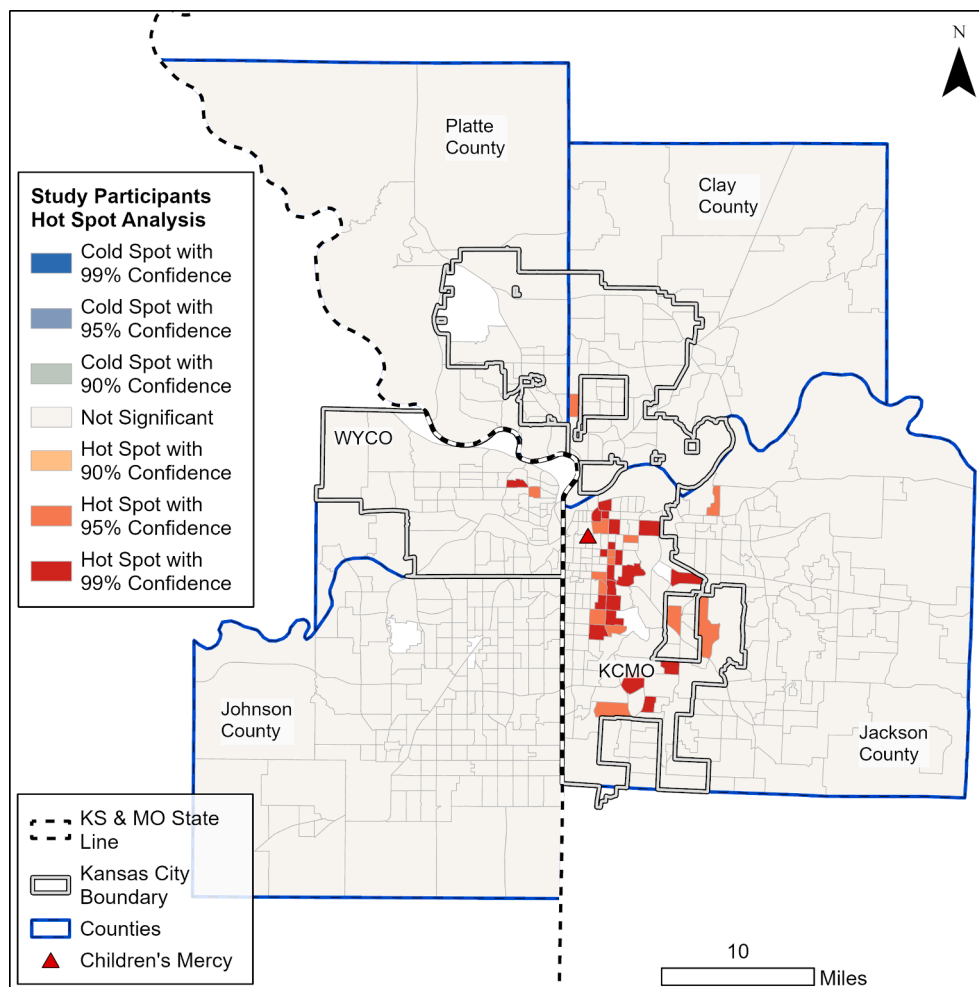
Fig. 3. Recruitment Hot Spots: Results of the Getis-Ord Gi* analysis on SEA recruitment by census tract. Results indicate census tracts in which a higher-than-expected count of SEA study participants were enrolled relative to nearby census tracts.

farther away from the matching geographic reference cohort than even a random sample of reference cohort. The attributes of the outliers (n = 6) whose empirical distance was above of the random distribution at 99% confidence ($\alpha < 0.01$, single sided) are presented in comparison to the mean attributes of the geographically matched reference cohort subgroup, in Table 3.

5. Discussion

Individuals enrolled in research often do not fully reflect the broader sociodemographic profile of the population that comprise the healthcare centers and labs performing these studies [31]. To understand then how best to generalize the results of a given study, we must continue efforts to understand the manner in which study cohorts do, or do not, accurately reflect the broader population from which they are drawn, not only in aggregate, but at the individual-level. Applying resources from medical informatics and data science methods can help researchers ensure that their cohort is as representative as possible.

Building on contemporary approaches for quantifying representativeness utilizing a reference cohort, this work introduces a novel approach to capture geospatial information linked to individuals' home addresses in such calculations. We recognize the impacts of housing and neighborhoods are complex and are unlikely to be appropriately captured or quantified by any single data point. Accordingly, this work was designed to focus not on the specific data elements that describe an individual, but on techniques to appropriately account for geographic

distribution in the assessment of a given metric.

This approach is a direct reflection of the diverse team structure of this project. From conception to completion over a dozen unique metrics were proposed, assessed, and refined. Most notably it was our community team members that raised the concern that use of geospatial data can delineate areas using arbitrary geospatial boundaries (e.g., census tracts), which may separate individuals near boundary lines that likely experience similar sociodemographic neighborhood influences. The global weighting approach was then developed to quantify alignment of two complete cohorts, an approach we believe to be novel in a representation comparison outside of statistical techniques around inverse probability weighting.

Looking at the case study, we find the ability of the proposed methods to quantify representativeness of the SEA study cohort relative to eligible but non-recruited individuals in the CMKC TSA highly encouraging. Turning first to the measures of recruitment, the joint assessment of geospatial distribution of the study and reference cohorts can provide unique and important insight into potential biases missed from viewing the cohorts as a single large group. The TSA is expansive, covering approximately 513 census tracts across two states. From a high-level review, we observe the refence cohort represents just under 75 % of the TSA, while the recruited population represents a significantly smaller percentage (29 %). Given the sample size of the study cohort relative to the reference cohort it is unsurprising we observe low absolute coverage (~39 %). We note, however, the study cohort has a well-aligned *expected* coverage rate (~80 %), suggesting these data are

Table 2

Balance of study and reference cohorts: Data are summarized across 500-bootstrap iterations of matching. Continuous variables are summarized with a Standardized Mean Difference (SMD), while nominal factors capture the largest proportion difference within levels of the respective factor. Panel (A) represents matching performed within the matching geography of each study patient. Panel (B) captures matching performed across the entire reference cohort, with matching weighted by Child Opportunity Index z-score.

Variable		Mean (SD)	[Min-Max]
(A) − Within Geography	*Age*	0.471 (0.294)	[0.003–1.651]
	Sex	0.049 (0.038)	[0–0.194]
	Insurance Type	0.058 (0.034)	[0.006–0.158]
	Ethnicity	0.036 (0.008)	[0.006–0.059]
	Acute Care Visits in Study Period	0.366 (0.093)	[0.047–0.6]
	Pediatric Comorbidity Index	0.52 (0.108)	[0.236–0.813]
	Distance to CMKC	0.018 (0.012)	[0–0.062]
(B) − Global Weighting	*Age*	1.386 (1.05)	[0.014–4.494]
	Sex	0.114 (0.089)	[0–0.428]
	Insurance Type	0.118 (0.071)	[0.012–0.428]
	Ethnicity	0.039 (0.015)	[0.006–0.22]
	Acute Care Visits in Study Period	0.418 (0.203)	[0.001–1.045]
	Pediatric Comorbidity Index	0.577 (0.295)	[0.003–1.696]
	Distance to CMKC	0.168 (0.123)	[0.002–0.754]
	Census Population Estimate	0.158 (0.117)	[0–0.556]
	Median Income	0.103 (0.077)	[0–0.395]
	Percent High School Education	0.179 (0.134)	[0.002–0.698]
	Percent of Vacant Houses	0.172 (0.132)	[0–0.658]

reasonably in-line with recruitment efforts for a study cohort of this size spread over such a wide geographic area. We expect study teams will use these metrics in tandem. Recognizing that recruitment is well-distributed for the allotted sample size, but also that analytical results of this sample may reflect of only a small area relative to where the broader population of individuals reside. This contrast is further supported by the misalignment in expected and absolute total land area. As the study cohort captured ~18 % fewer sq. miles than expected from the 1000 + sq. miles of the reference cohort.

These values can be further contextualized with measures of recruitment rate. Alignment of absolute and expected recruitment (~13 %) indicate low but anticipated rates in a given census tract. Again, we expect study teams to jointly utilize metrics to adapt recruiting and analysis plans. From these data it is clear recruitment has gone well, but due to sample size, is expected to capture only a fractional percentage of eligible patients in any given area. Thus, it will be critical to assure the subset of study members are representative of the reference cohort in any given area (*Individual Characteristic Metrics*). Such information can provide insight into generalization of study results, as imbalanced recruitment alone does not inherently mean that the cohort is fundamentally biased. Hotspot analysis provides additional information, in this case illuminating a concerning pattern of clustered recruitment. Even with balanced recruitment by percentage of eligible patients, this finding suggests it may warrant targeted efforts to broaden recruitment

to areas with smaller pools of candidate individuals.

Moving to the assessment of alignment on individual-level characteristics, we note several benefits in the consideration of geographic data between study and reference cohorts. First, evaluating alignment through repeated case-control matching, we can identify the subset of features for which the study cohort closely resembles the broader reference cohort, and those features for which they significantly deviate. This information can be used to improve recruitment and analysis. On the front end, recruitment can be prospectively monitored and adjusted to align factors known to be confounders in downstream analysis. While at the time of analysis, study teams can justify the utilization of appropriate techniques (e.g. doubly robust regression) to better account for unbalanced factors.

Finally, distance-based comparisons provide an opportunity to dive deeper into misalignment, identifying specific individuals in the study cohort that are in some way distinct from the reference cohort on which the results may be expected to generalize. For example, from Table 3, the age of and count of acute care visits for patients 1 and 3, or the pediatric severity score of patient 4. This information offers study teams a quantitative measure to justify potential exclusion or stratification in analysis as a means to better align study results with broader reference community.

As both approaches to individual-level alignment produce standardized continuous values, a study team is able to determine a threshold of alignment based on study parameters and published literature. Teams also have an option to consider alignment within-geographies or across the full reference cohort. While comparisons within the same geographic area offer a more direct assessment, global metrics can measure alignment in regions with limited reference cohorts or for highly heterogenous study cohorts. However, the quality of this comparison is highly dependent on the weighting factor used, which must be closely related to similarity between individuals. Additionally, even with weighted sampling, there is expected decrease in alignment score using global metrics, as samples (infrequently) include disparate individuals. Suggesting these metrics may require additional sampling iterations for robustness.

5.1. Considerations for use

While alignment on geospatial factors is perhaps most commonly associated with public-health studies, the relationship between location and individual-level health suggests we must consider these measures for all types of research. Yet as researchers strive for cohort alignment and generalizability, it is important to recognize the determination of *which* measures require balance remains a key question in study design. Alignment should be viewed as a holistic concept, and not every factor must be perfectly balanced for every study. A combination of both the study cohort and broader research question should guide how imbalance on specific metrics is be considered and addressed. For example, given that the SEA study focuses on asthma and there are known associations between asthma exacerbation and home proximity to major roadways, it is perhaps more important to balance measures of coverage than spread, assuring we capture a representative set of geographic regions (rather than simply a large area) in which eligible individuals reside.

Furthermore, development of a temporally aligned reference cohort is also critical. By comparing SEA cohort against individuals seen (but not recruited) during the SEA recruitment period, we align changing sociodemographic environments across the geographies being compared, an important consideration for all studies utilizing these methods. In making such decisions, we advocate for collaboration with individuals in the communities being evaluated, as their insights often capture perspectives on study design not available from quantitative data alone. This information can be complemented with a computational approach, where determination of impactful factors can be measured across geospatial metrics a-priori using the reference cohort alone. By drawing samples of the expected study cohort size from the

Table 3
Overview Distance-Based Outliers: This table outlines each of the 6 identified outliers in the measures of individual characteristics. For each case, checkmarks in row (P) specifies in the data for the outlier, while (R) provides the distribution of each factor in the reference cohort.

		Age	Sex		Ethnicity		Insurance Type				Acute Care Visits in Study Period	Pediatric Comorbidity Index	Distance to CMKC
			Female	Male	Hispanic	Non-Hispanic/Non-Latino	Commercial	Commercial & Medicaid	Medicaid	Self-Pay			
1	P	0.25		✓		✓	✓				15.00	0.29	11.30
	R	9.40 (5.05)	5 (50.00)	5 (50.00)	0 (0.00)	10 (100.00)	1 (10.00)	3 (30.00)	6 (60.00)	0 (0.00)	3.00 (3.37)	0.91 (0.27)	11.52 (0.38)
2	P	5.92	✓			✓			✓		4.00	0.40	5.22
	R	9.81 (6.03)	4 (44.44)	5 (55.56)	0 (0.00)	9 (100.00)	0 (0.00)	2 (22.22)	6 (66.67)	1 (11.11)	2.00 (0.87)	0.85 (0.29)	5.20 (0.33)
3	P	0.92		✓		✓	✓				12.00	0.19	4.53
	R	8.40 (5.41)	13 (52.00)	12 (48.00)	0 (0.00)	25 (100.00)	2 (8.00)	0 (0.00)	22 (88.00)	1 (4.00)	2.80 (2.12)	0.89 (0.37)	4.17 (0.28)
4	P	3.17		✓	✓			✓			6.00	0.31	14.93
	R	7.92 (2.60)	2 (50.00)	2 (50.00)	0 (0.00)	4 (100.00)	2 (50.00)	0 (0)	2 (50.00)	0 (0.00)	2.00 (0.82)	1.17 (0.33)	15.65 (0.02)
5	P	2.50				✓			✓		7.00	0.33	18.81
	R	8.76 (4.78)	4 (23.53)	13 (76.47)	0 (0.00)	17 (100.00)	5 (29.41)	2 (11.76)	10 (58.82)	0 (0.00)	2.53 (1.55)	0.63 (0.36)	20.35 (0.19)
6	P	5.00		✓	✓				✓		6.00	0.33	12.47
	R	6.42 (4.33)	12 (52.17)	11 (47.83)	2 (8.70)	21 (91.30)	2 (8.70)	0 (0.00)	21 (91.30)	0 (0.00)	2.87 (2.16)	0.70 (0.34)	11.68 (0.47)

pool of reference individuals (as best determined from available retrospective data), it is possible to assess factors for which variability is expected even before recruiting the first individual. During recruitment these factors can be prospectively monitored, and recruitment targeted to achieve balance.

5.2. Limitations

Consideration of geospatial information presents several challenges. First, in addition to typical concerns of data quality (ambiguous addresses, etc.) [32], individual addresses may no longer reflect an individual's current living situation and thus bias estimates, particularly for those experiencing housing insecurity. These concerns can be partially mitigated by regular data collection inherently collected from larger health systems that provide comprehensive emergency, primary care, and/or urgent care services but may require careful curation for small study centers or studies involving vulnerable populations. Second, the addition of a geographical context to alignment measures inherently represents additional level of data stratification. As a result, for highly specific study cohorts, reference cohorts for exact geographies may become too sparse to obtain reliable estimates. We have taken steps to address this by introducing measures that utilize data across the entire reference cohort, and the utilization of expected measures based on sample size for the assessment of recruitment. However, it remains important to consider the granularity at which geocoding takes places (highly specific census block group, vs larger regions such as counties) based on the scope of the study and intended generalization of results.

Finally, this approach requires the same data elements be available for both the study cohort and reference cohorts. While this may be possible for large healthcare systems, with large repositories of historical data, this may not be always possible for other institutions, especially if specialized data is collected to recruit the study cohort. In a similar vein, we note use of historical records can introduce bias into reference cohort selection. In this case study we noted not all members of the SEA

study could be identified using the inclusion criteria. Post-hoc investigation found some had EHR race data defined as multiracial, but who identified as Black/African American on recruitment. Although it does not impact the results of the presented case study, as our reference cohort only considers non-recruited individuals, this is an important reminder that any comparison of cohorts can be biased by local coding and documentation practices.

6. Conclusion and future development

Inequity in research recruitment can have lasting impacts on future research by excluding certain populations, including many underserved groups. Whether it be improved understanding of the limitations of existing data or a means to monitor and adjust prospective data collection, the informatics methods presented in this paper illustrate clear value in capturing geographic context in the assessment of cohort representativeness. While it is important to remember challenges in generalization may stem from elements of study design, the ability to capture and describe heterogeneity in the cohort remains an often overlooked responsibility of a study team for which limited guidance is provided [33]. As the secondary-use of data has become a key element in training of large computational models, the ability to account for imbalances relative to the general population will be critical in mitigating the proliferation of biased and potentially harmful systems. It is our hope this work provides a foundation on which to best make use of the insights gained from research studies by improving the ability to understand the characteristics of a cohort best suited to generalize results.

Declaration of generative AI and AI-assisted technologies in the writing process

None.

Data availability

The data used in this analysis were aggregated from operational clinical sources with the approval of the Children's Mercy Kansas City Review Board (CMKC IRB) and due to concerns of patient identification are not posted publicly. The data may be shared on reasonable request to the corresponding author but note such requests may need separate approval by the CMKC IRB.

CRediT authorship contribution statement

Keith Feldman: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **Natalie J. Kane:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Stacey Daniels-Young:** Writing – review & editing, Validation, Methodology. **Bruce Reed:** Writing – review & editing, Validation, Methodology. **Jessica Welch:** Writing – review & editing, Validation, Methodology, Conceptualization. **Lauren Fitzpatrick:** Writing – review & editing, Software. **Mark A. Hoffman:** Writing – review & editing, Resources, Methodology, Conceptualization. **Andrea Bradley-Ewing:** Writing – review & editing, Validation, Methodology, Conceptualization. **Elin Grundberg:** Writing – review & editing, Validation, Resources, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2024.104687.

References

[1] Services DoHaH, Protection of Human Subjects, 45 C.F.R. §46.102, 2018.
[2] W.A. Kukull, M. Ganguli, Generalizability: the trees, the forest, and the low-hanging fruit, Neurology. 78 (23) (2012) 1886–1891.
[3] D.F. Polit, C.T. Beck, Generalization in quantitative and qualitative research: myths and strategies, Int. J. Nurs. Stud. 47 (11) (2010) 1451–1458.
[4] A.M. Briesch, H. Swaminathan, M. Welsh, S.M. Chafouleas, Generalizability theory: a practical guide to study design, implementation, and interpretation, J. School Psychol. 52 (1) (2014) 13–35.
[5] A.J. Averitt, P.B. Ryan, C. Weng, A. Perotte, A conceptual framework for external validity, J. Biomed. Inform. 121 (2021) 103870.
[6] A.G. Huebschmann, I.M. Leavitt, R.E. Glasgow, Making health research matter: a call to increase attention to external validity, Ann. Rev. Public Health. 40 (2019) 45–63.
[7] E.A. Stuart, C.P. Bradshaw, P.J. Leaf, Assessing the generalizability of randomized trial results to target populations, Prevent. Sci. 16 (2015) 475–485.
[8] P.M. Rothwell, Factors that can affect the external validity of randomised controlled trials, PLoS Clin. Trials. 1 (1) (2006) e9.
[9] J.E. Rudolph, Y. Zhong, P. Duggal, S.H. Mehta, B. Lau, Defining representativeness of study samples in medical and population health research, BMJ Med. 2 (1) (2023).
[10] S. Li, D.F. Heitjan, Generalizing clinical trial results to a target population, Stat. Biopharm. Res. 15 (1) (2023) 125–132.
[11] E. Tipton, Improving generalizations from experiments using propensity score subclassification: assumptions, properties, and contexts, J. Educ. Behav. Stat. 38 (3) (2013) 239–266.
[12] Y. Liu, A. Gelman, Q. Chen, Inference from nonrandom samples using bayesian machine learning, J. Surv. Stat. Methodol. 11 (2) (2023) 433–455.
[13] A. Coppock, T.J. Leeper, K.J. Mullinix, Generalizability of heterogeneous treatment effect estimates across samples, Proc. Natl. Acad. Sci. 115 (49) (2018) 12441–12446.
[14] X. Wang, S. Piantadosi, J. Le-Rademacher, S.J. Mandrekar, Statistical considerations for subgroup analyses, J. Thor. Oncol. 16 (3) (2021) 375–380.
[15] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, A survey on bias and fairness in machine learning, ACM Comput. Surv. (CSUR). 54 (6) (2021) 1–35.
[16] S. Caton, C. Haas, Fairness in machine learning: a survey, ACM Comput. Surv. (2020).
[17] Z. He, X. Tang, X. Yang, et al., Clinical trial generalizability assessment in the big data era: a review, Clin. Transl. Sci. 13 (4) (2020) 675–684.
[18] A. Getis, J.K. Ord, The analysis of spatial association by use of distance statistics, Geogr. Anal. 24 (3) (1992) 189–206.
[19] C. Redlands, Environmental Systems Research Institute, Inc(esri) (2004).
[20] J.C. Gower, A general coefficient of similarity and some of its properties, Biometrics. 857–871 (1971).
[21] P.C. Austin, Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples, Stat. Med. 28 (25) (2009) 3083–3107.
[22] E.A. Stuart, B.K. Lee, F.P. Leacy, Prognostic score–based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research, J. Clin. Epidemiol. 66 (8) (2013) S84–S90.
[23] S.V. Belitser, E.P. Martens, W.R. Pestman, R.H. Groenwold, A. De Boer, O.H. Klungel, Measuring balance and model selection in propensity score methods, Pharmacoepidemiol. Drug Safety. 20 (11) (2011) 1115–1129.
[24] P.L. Remington, B.B. Catlin, K.P. Gennuso, The county health rankings: rationale and methods, Population Health Metrics. 13 (1) (2015) 1–12.
[25] The State of Children's Health, 2019 Community Health Needs Assessment for the Kansas City Region, Children's Mercy Kansas City, Kansas City, MO, 2019.
[26] C. Brokamp, C. Wolfe, T. Lingren, J. Harley, P. Ryan, Decentralized and reproducible geocoding and characterization of community and environmental exposures for multisite studies, J. Am. Med. Inform. Assoc. 25 (3) (2018) 309–314.
[27] J.W. Sun, F.T. Bourgeois, S. Haneuse, et al., Development and validation of a pediatric comorbidity index, Am. J. Epidemiol. 190 (5) (2021) 918–927.
[28] K. Walker, H.M. Tidycensus, Load us census boundary and attribute data as' tidyverse'and'sf'-ready data frames, R Package Vers. 1 (2021).
[29] Team R, RStudio: integrated development for R. (No Title), 2015.
[30] P. Virtanen, R. Gommers, T.E. Oliphant, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods. 17 (3) (2020) 261–272.
[31] Editorsm, Striving for diversity in research studies, in: Vol 385: Mass Med. Soc. (2021) 1429–1430.
[32] C. Owusu, Y. Lan, M. Zheng, W. Tang, E. Delmelle, Geocoding fundamentals and associated challenges, Geos, Data Sci. Tech. Appl. 118 (2017) 41–62.
[33] J. Moeller, J. Dietrich, A. Neubauer et al., Generalizability crisis meets heterogeneity revolution: determining under which boundary conditions findings replicate and generalize, 2022.