Breakout Session 6: Track A

Consideration of Geospatial Distribution in the Measurement of Study Cohort Representativeness and Data Quality

Dr. Keith Feldman (Moderator)
Assistant Professor, Children's Mercy Kansas City

Contextualizing and Addressing Population-Level Bias in a Social Epigenomics Study of Asthma

Supplement R01MD015409-03S1

Speaker: Keith Feldman, PhD, PI: Elin Grundberg PhD

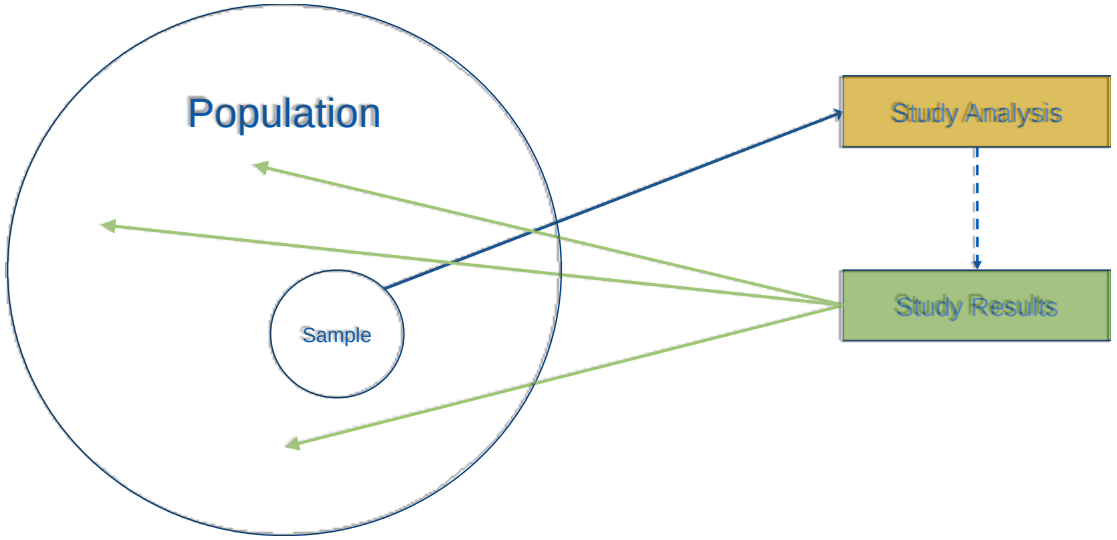
UMKC School of Medicine

The Problem with Generalization

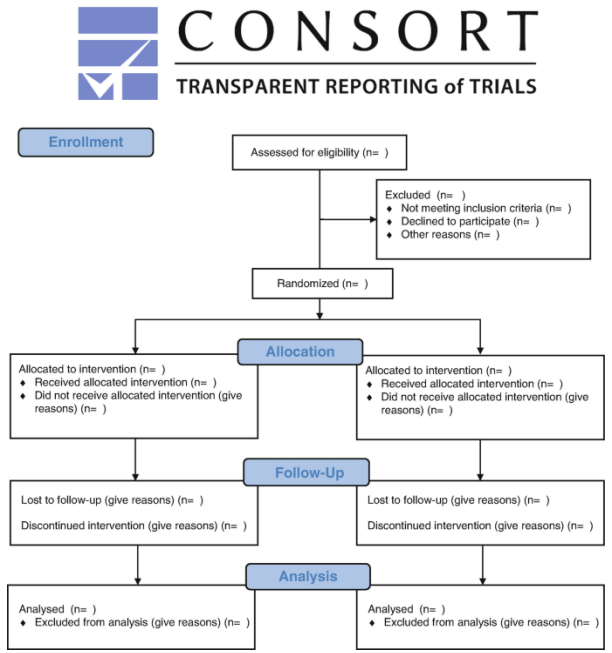
Study results are a product of the data (subjects) used for analysis

Traditional "Table 1"

		Study
	n	380
Demographics	**Age (Years)**	6.21 (4.35)
	Sex	
	Female	173 (45.53)
	Male	207 (54.47)
	Insurance Type	
	Commercial	45 (11.84)
	Commercial & Medicaid	23 (6.05)
	Medicaid	303 (79.74)
	Self-Pay	8 (2.11)
	Unknown	1 (0.26)
	Ethnicity	
	Hispanic/Latino	14 (3.68)
	Non-Hispanic/Non-Latino	364 (95.79)
	Unknown	2 (0.53)

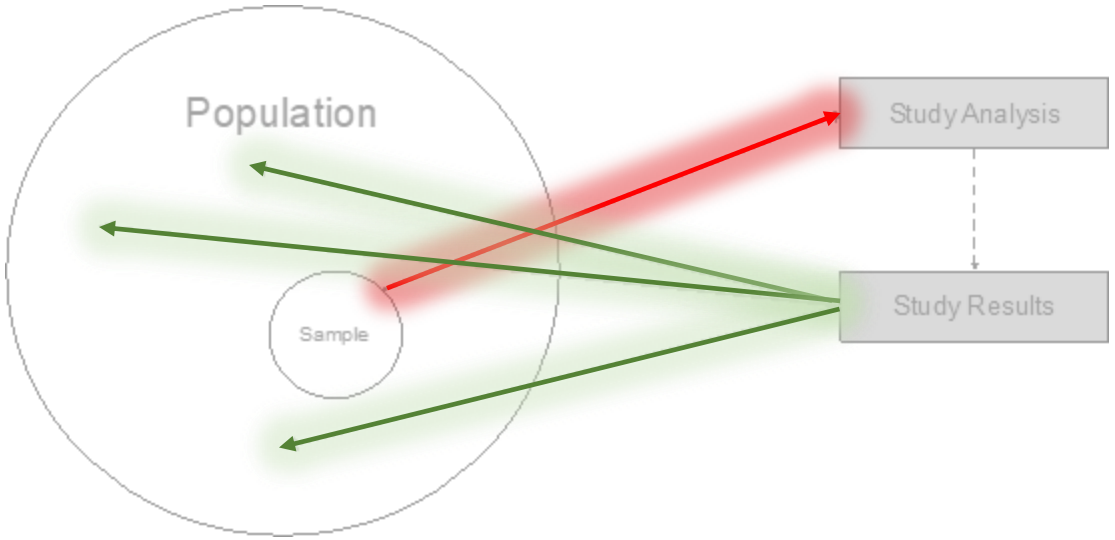


Adapted From: Kamper, Steven J. "Generalizability: linking evidence to practice." journal of orthopaedic & sports physical therapy 50.1 (2020): 45-46.



Project Objectives

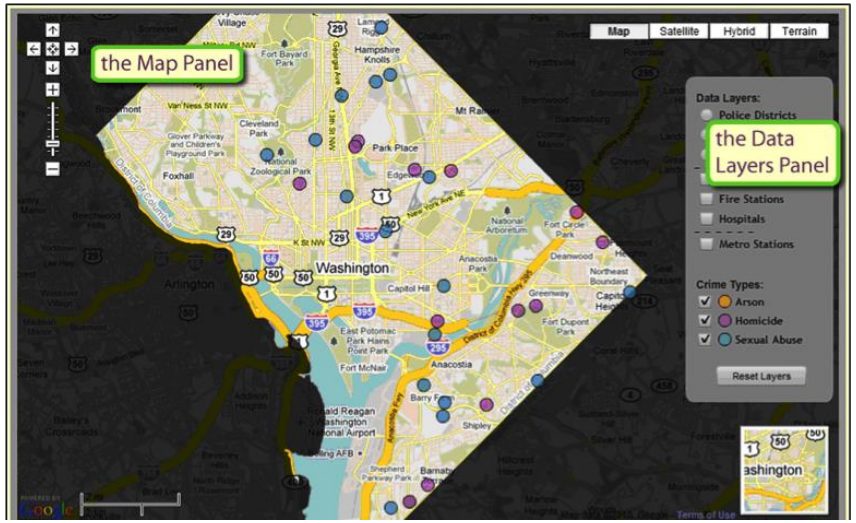
Develop methods to quantify the representativeness of a study sample against a reference cohort to better understand for whom the study results can be expected to generalize, and to develop methods robust to such imbalances



Quantify Representativeness

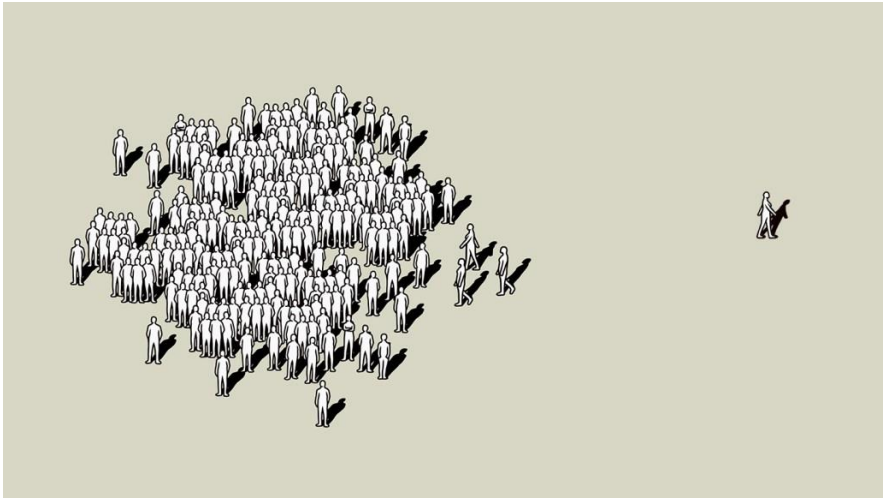
Geospatial Representation

Benchmarking characteristics of study cohort against a reference population on whom you want to assess potential generalizability.



Data Quality

Quantify patterns of data quality across subgroups of data and over time. Assess impact of data quality issues on reliability / stability of study results



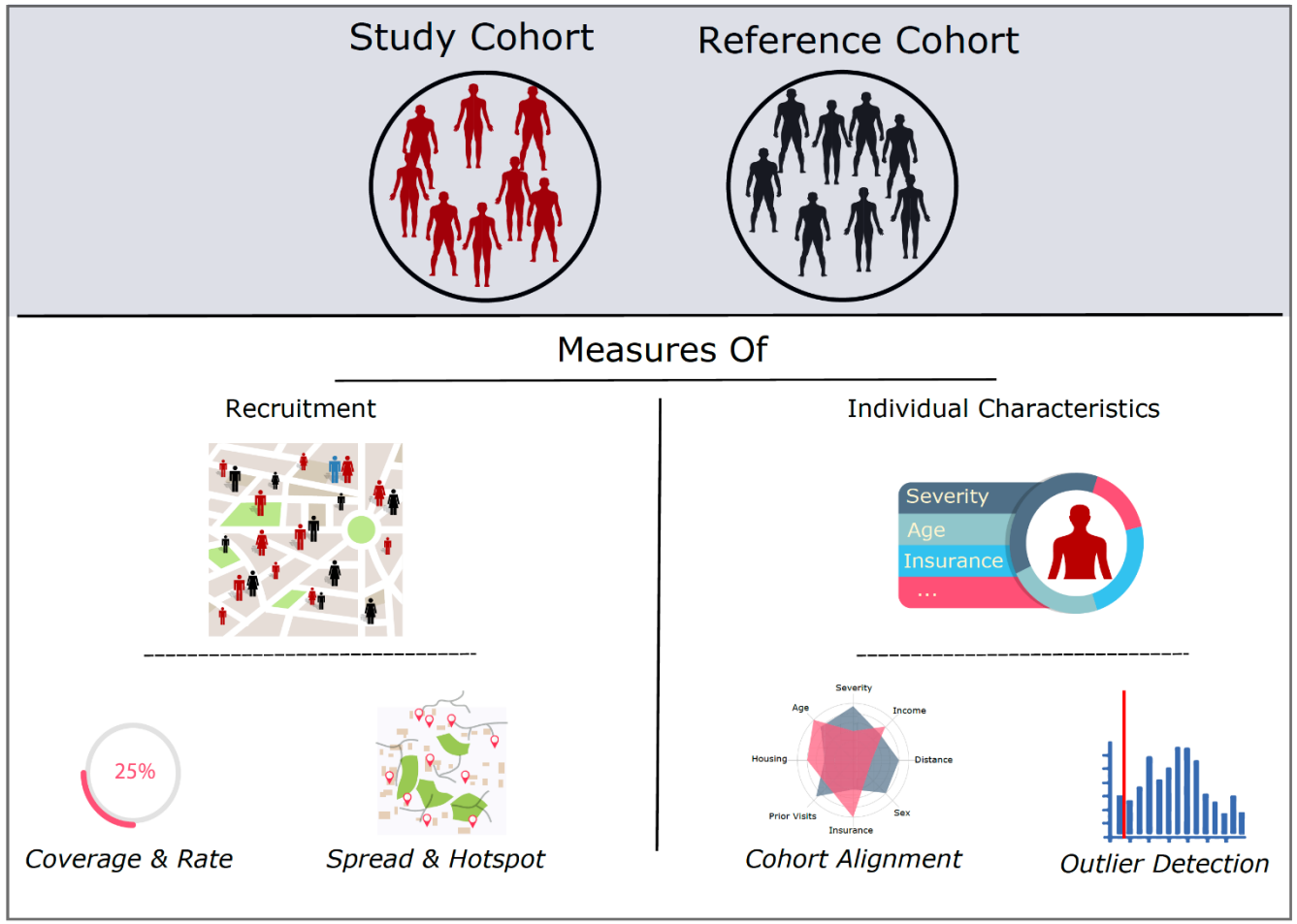
Parent Award Cohort and Data

The Stress, Epigenome and Asthma (SEA) study focuses on disentangling mechanisms by which exposure to chronic stress may cause epigenetic changes increasing susceptibility to rhinoviral (RV) infection – and thus increased risk of asthma exacerbations

- Study Cohort
  - SEA study (03/15/2021 -11/08/2022) - 400 English-speaking Black/African American children, age <18 presenting to ER with acute respiratory symptoms related to an asthma diagnosis
- Reference Cohort (same period)
  - 12,699 Black/African American children with a history of Asthma.
  - 2,757 children were identified as having at least 1 encounter for an Asthma exacerbation during the study period.

- **Location:**
  - Distance between the individual's address and recruitment location (CMKC Adelle Hall)
- **Socio-Demographics:**
  - Age at encounter, primary insurance, sex assigned at birth, as well as self-reported race, ethnicity, and nationality.
- **Clinical :**
  - Pediatric comorbidity index for all visits during the study period and derived count of eligible visits during the study.
- **Population-Level Indicator Data:**
  - Fraction of population >25 with educational attainment of at least high school graduation, and fraction of houses that are vacant.

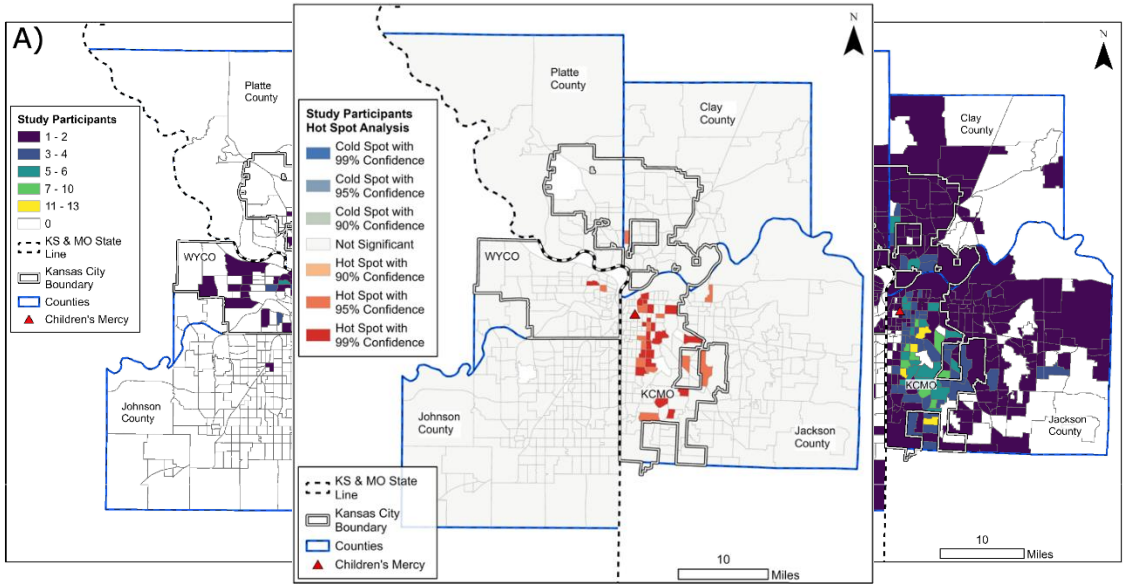
Geospatial Assessment



Measures of Recruitment

Coverage and Recruitment	Coverage: Number of distinct regions of recruited subjects are drawn from compared to number of possible regions for recruitment Can be calculated at any granularity of geospatial geography (Counties, Census Block groups, etc.) as defined by the user.
	Recruitment Rate: Percentage eligible patients in each geographic area that are recruited to the study. This value can then be aggregated across the set of geographic areas in the reference set to summarize recruitment averages and variability.
Spread and Hot Spots	Spread: The total geographic area across the unique geographies that comprise the study cohort. To identify situations in which high-recruitment represents a limited geographic area
	Hot Spots: Measure of geographic clustering to identify an area with a higher concentration of recruited subjects compared to the expected number given a random distribution of subjects.

Recruitment Hot Spots: Results of the Getis-Ord Gi* analysis of SEA study cohort by census tract. Red areas indicate tracts in which each of the 380 expected counts of SEA study participants were enrolled at Children's Mercy Kansas City



Coverage of approximately 39.1% (149 Tracks Study Cohort, 380 Tracks Reference).

Spread: Although a moderately sized absolute number, the 149 unique tracts represent a relatively small geographic area of just 234.47 sq. miles. This represents only about 23% of the total coverage of the total area reference cohort (1000.27 sq. miles).

Measures of Recruitment

Case Control	*Create an aggregate measure of how well aligned the study and reference cohorts are given a set of factors* Measures deviation between the study and reference populations as "balance" in case-control matches. For each study cohort members, 1 random individual is drawn from the reference cohort in the same geographic area. Pairs are treated as "matches" and can directly leverage an array of established case-control balance metrics, including measures of standardized mean difference (continuous), and maximal proportion difference (nominal) data
Distance Based	*Identify Individual-level Measures of Similarity + Outlier Detection* For each member of the study cohort, distance is computed against all individuals from of the reference population drawn who resides in the same geographic region. Used to compute a measure of how likely the average distance between study member and reference cohort is to happen by chance. Used to compute a measure of the expected intra-individual distance within the reference cohort. How close is the distance between study member and average reference distance.

Overview Distance-Based Outliers: *This table outlines each of the 6 identified outliers in the measures of individual characteristics. For each case,* ***Balance of study cohort and reference population:*** *500-bootstrap iterations of matching. Continuous variables are summarized with a Standardized Mean Difference (SMD), while nominal factors capture the largest proportion checkmarks in row (P) specifies in the data for the outlier, while (R) provides the distribution of each factor in the reference population difference within levels of the respective factor. Panel (A) represents matching*

			Sex		*Ethnicity*		*Insurance Type*				
		Age	Female	Male	Hispanic	Non-Hispanic/ Non-Latino	Commercial	Commercial & Medicaid	Medicaid	Self-Pay	Acute Care Visits in Study Period
1	*P*	0.25		✓		✓	✓				15.00
	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)
2	*P*	5.92	✓			✓			✓		4.00
	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)
3	*P*	0.92		✓		✓	✓				12.00
	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)
4	*P*	3.17		✓	✓			✓			6.00
	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)
5	*P*	2.50		✓		✓			✓		7.00
	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)
6	*P*	5.00		✓	✓				✓		6.00
	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)

Data Quality

Assume we have 3 variables: Age, Sex, Recruitment Location

Sex	Male	
	Female	
Age (percentile)	0-25th	25th – 50th
	50th – 75th	75th – 100th
Recruitment Location	Inpatient (1)	
	ED (2)	

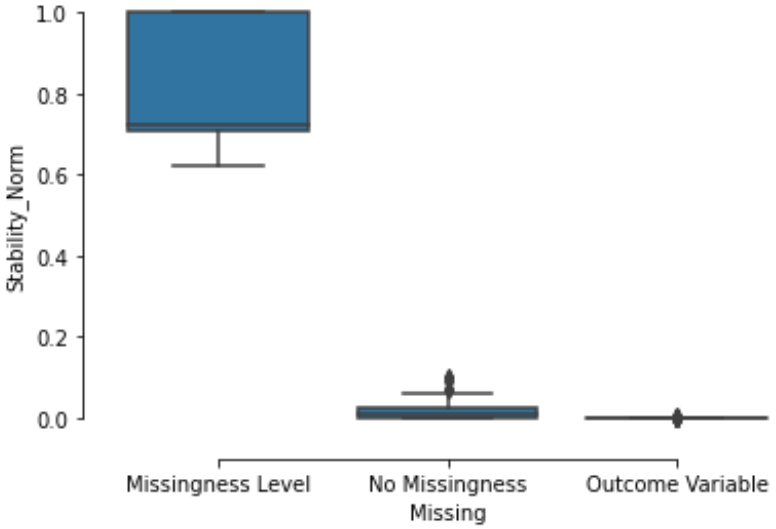
Goal 1: Identify % of data in various subgroups

Configuration	Group_Val	Num_Params	Size	Val_AgeQ_1	Val_AgeQ_2	Val_AgeQ_3	Val_AgeQ_4
Female	(Sex,)	1	135	0.207407	0.244444	0.229630	0.318519
Male	(Sex,)	1	165	0.284848	0.254545	0.266667	0.193939
1	(Location,)	1	265	0.252830	0.264151	0.241509	0.241509
2	(Location,)	1	35	0.228571	0.142857	0.314286	0.314286
('Female', 1)	(Sex, Location)	2	120	0.216667	0.241667	0.225000	0.316667
('Female', 2)	(Sex, Location)	2	15	0.133333	0.266667	0.266667	0.333333
('Male', 1)	(Sex, Location)	2	145	0.282759	0.282759	0.255172	0.179310
('Male', 2)	(Sex, Location)	2	20	0.300000	0.050000	0.350000	0.300000

Subgroup Bias Metric

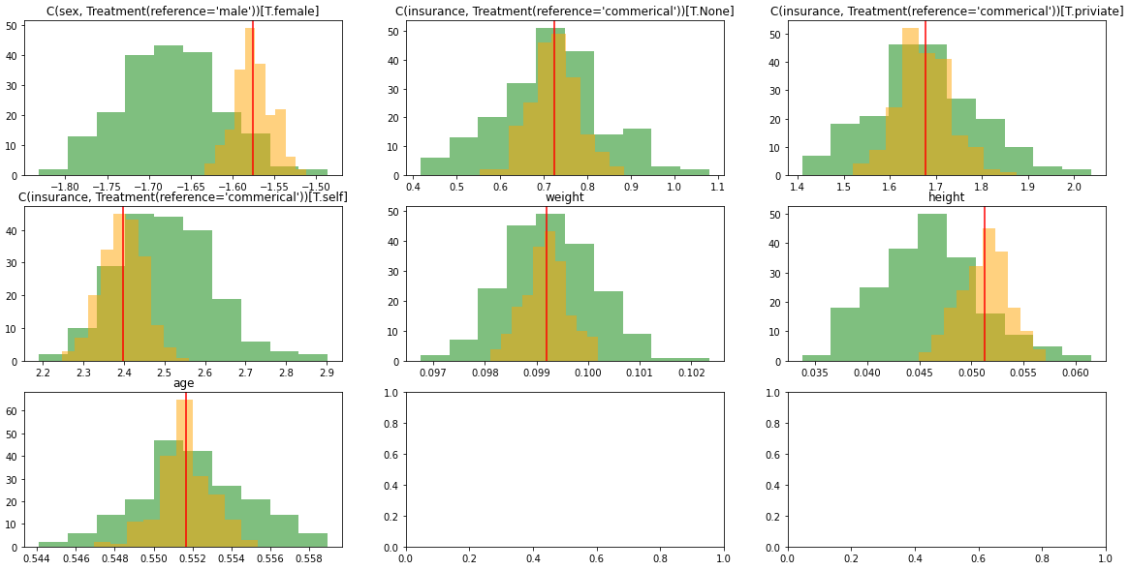
Aggregate Measure

Developed a series of metrics to quantify the degree to which missingness for a given subspace significantly differs from levels of the factors that comprise it.



Impact Analysis

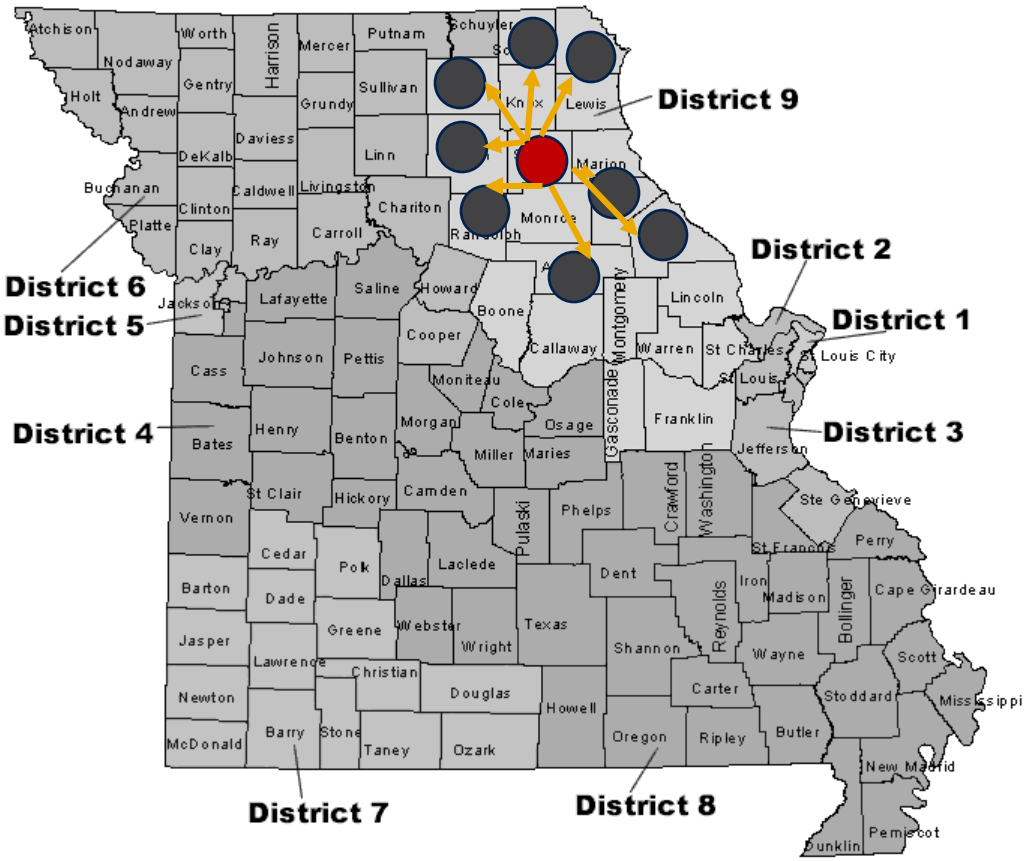
Verify if missingness in sub-groups results in biased estimates for downstream analysis, or if wider confidence intervals will account for this



Geospatial Boundaries

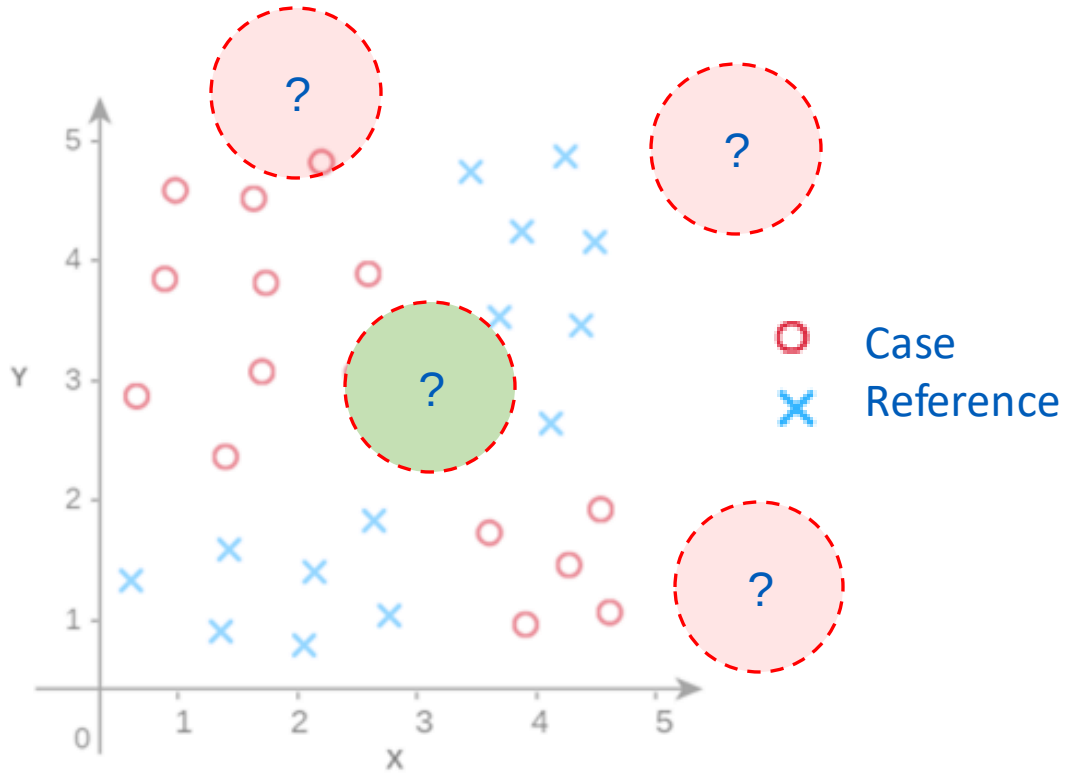
- Comparing individuals within exactly matching geographic regions draws an arbitrary delineation in space, where subjects only a small absolute distance apart may be separated by a census block line and are thus not considered in the alignment measure.

- We introduce extensions of the previously described methods that leverage the entire reference population. This provides two benefits:
  - Allows for improved estimates of representations for study cohort in regions with small reference populations.
  - Second, it allows for the inclusion of data captured at an aggregate/population level (e.g., average access to mental Health Providers).



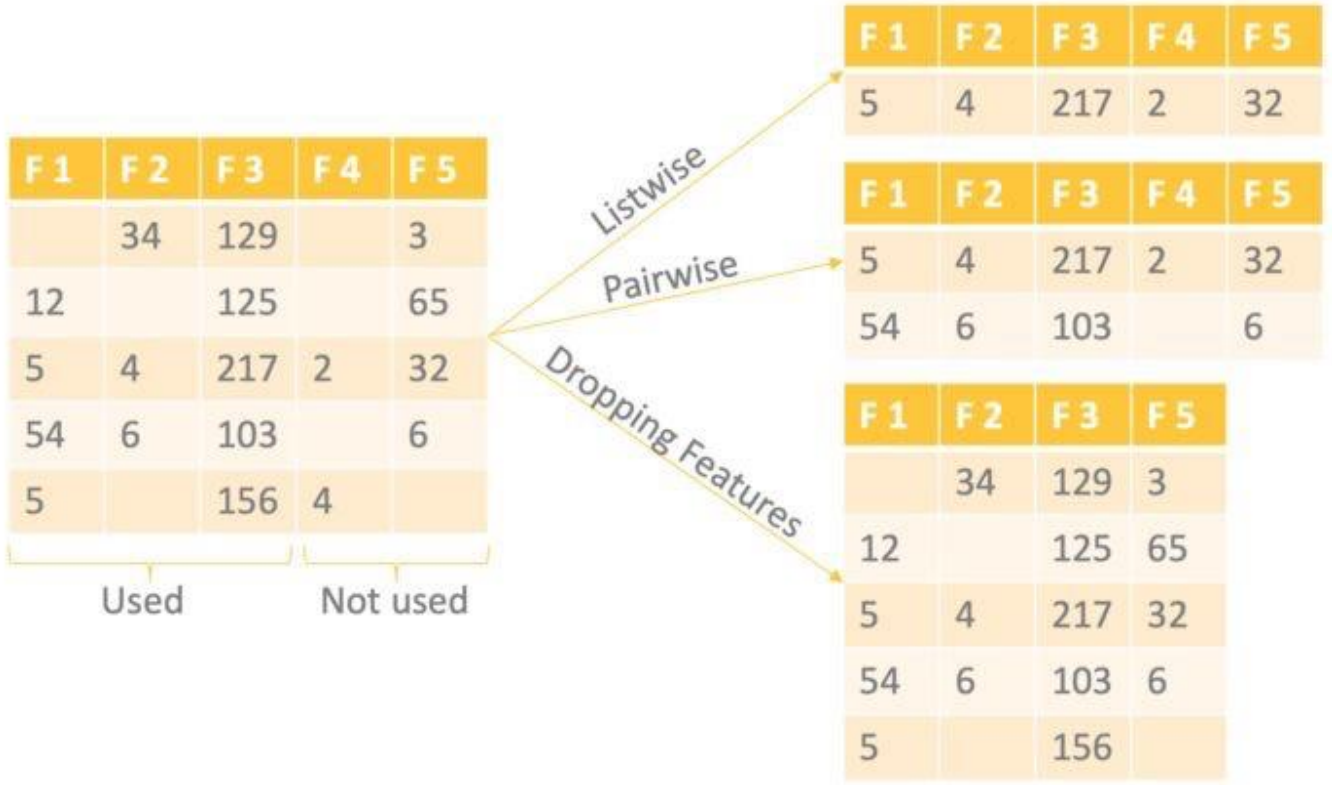
Extrapolation vs Out of Reference

Does the Data Represent The Full Range of Expected Values



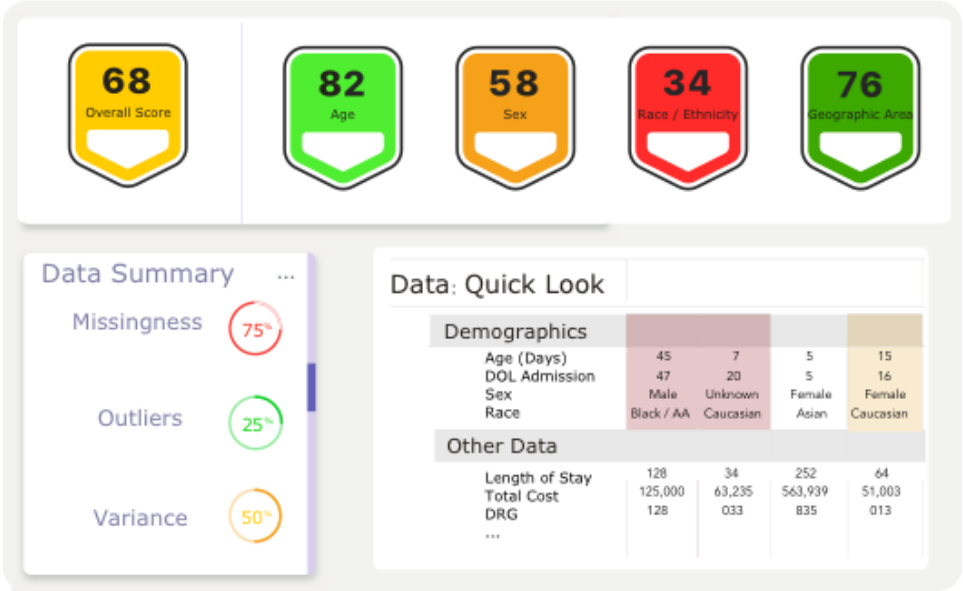
Imputation and Bias – Second Aim

In Real-World Studies, compete data of any kind is rare



Ongoing Work

Standardized metrics pertaining to the study cohort representativeness and quality will help inform researchers of inherent biases, limits of conclusions and more broadly improve generalizability for future cohort studies, allowing for safer reuse of data



Integration into CTSI Informatics Core

FRONTIERS
CLINICAL & TRANSLATIONAL SCIENCE INSTITUTE
AT THE UNIVERSITY OF KANSAS

Children's Mercy KANSAS CITY · KU · UMKC · Saint Luke's · Kansas City University · Kansas State University

As part of the informatics core of our regional CTSI program, geospatial measurement tools are being created. Working with this core, led by CO-I Dr. Mark Hoffman, we hope to integrate these measures into the developed toolset for researchers across the region