Workshop: Advanced Quantitative Methods for Innovative HIV Research

01/23/2023

Biostatistics and Computational Resources Team (BCRT)

Patrick Janulis, PhD
BCRT Co-Lead
Sr. Biostatistician

Anna Hotton, PhD
BCRT Co-Lead
Sr. Epidemiologist

Christina Hayford MS/MSP
Sr. Research Data Analyst

Susheel Reddy, MPH
Biostatistician

Angel Aviña, MS, MPH
Core Manager

Pre-Award Services

- Basic Study Design
- Analysis Plan
- Power Calculations
- Survey Methodology
- Geospatial Sampling
- Maps for Grant Applications

Basic Study Design

- Proper characterization of research objectives prior to initiation of analysis.
- Review data management paradigm prior to data collection
- Identify problematic issues relevant to inference
  - Violation of assumptions
  - Identifying/Accommodating missingness
  - Need for sensitivity or ancillary analyses

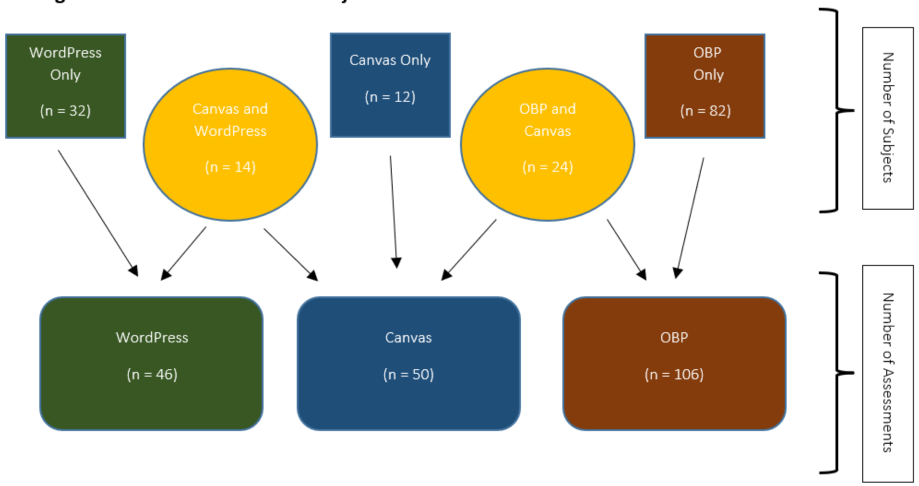
Analysis Plan

- Written statistical analysis plan
  - Clearly identifies for the investigator the analyses to be done, and objectives to be fulfilled.
  - Informs subsequent write-up and/or presentation
  - Valuable resource for analysis implementation.
- Can include analysis dataset specifications and display specifications (e.g., table shells)

Analysis Plan

- Can include analysis dataset specifications and display specifications (e.g., table shells)

Figure 1: Planned Number of Subjects and Assessments in Matched Cohort



4 STATISTICAL ANALYSIS

4.1 Patient Disposition

A table indicating the number and percentage of subjects who performed an assessment will be provided as an overview of the cohort disposition using the FAS. The number and percentage of assessments performed on each platform will be provided as well. Counts of number of subjects with one assessment and number of subjects with two assessments by platform will be included. This table will be similar, if not identical to Table 10 shown above (see Section 2).

In addition, the final analysis will include a figurative representation of the matched cohort data used in the primary analysis that will specify both the number of subjects and assessments used for inferential analyses applied to the MC. The figure will be similar, if not identical, to Figure 1 shown above.

Dataset Name: MCANLY
Dataset Description: Analysis dataset that represents all observations from FASANLY for subjects that are identified as being part of the matched cohort.
Program Location: P:\ID\SMART\Acceptability Testing\programs\dataset\mcanly.sas
Source Datasets: FASANLY, PSSANLY
Dataset Structure: One record per s
identified as being part of the prop

- 46 observations where PL
  WordPress)
- 50 observations where PL
  and 24 for subjects that c
- 106 observations where P
  Canvas)
- Total n = 202 for 164 subj

Identifiers: SMARTID (SOURCE.SMA
Sort Order: MATCHID, SMARTID, PL
General Notes:
To create MCANLY

1. Start with PSSANLY.
2. Create an output dataset
3. Keep only the propensity
4. Merge this working datas
5. Drop all observations whe

- Variables specified below
- See the Statistical Analysi

Table TBA: Summary of Least Square Mean Acceptability Score Estimates Using Mixed Model
Population: Matched Cohort

		OBP (N = 106)	Canvas (N = 50)	WordPress (N = 46)	P-value	Pairwise Differences OBP - Canvas	P-value	OBP - WordPress	P-value
AT Composite Scores [1]									
Overall	Number of Assessments	xx	xx	xx	x.xxxx		x.xxxx		x.xxxx
	LS Mean Estimate (SE)	x.x (x.xx)	x.x (x.xx)	x.x (x.xx)		x.x (x.xx)		x.x (x.xx)	
	95% CI	(x.x, x.x)	(x.x, x.x)	(x.x, x.x)		(x.x, x.x)		(x.x, x.x)	
Impact	Number of Assessments	xx	xx	xx	x.xxxx		x.xxxx		x.xxxx
	LS Mean Estimate (SE)	x.x (x.xx)	x.x (x.xx)	x.x (x.xx)		x.x (x.xx)		x.x (x.xx)	
	95% CI	(x.x, x.x)	(x.x, x.x)	(x.x, x.x)		(x.x, x.x)		(x.x, x.x)	
Usefulness	Number of Assessments	xx	xx	xx	x.xxxx		x.xxxx		x.xxxx
	LS Mean Estimate (SE)	x.x (x.xx)	x.x (x.xx)	x.x (x.xx)		x.x (x.xx)		x.x (x.xx)	
	95% CI	(x.x, x.x)	(x.x, x.x)	(x.x, x.x)		(x.x, x.x)		(x.x, x.x)	
Engagement	Number of Assessments	xx	xx	xx	x.xxxx		x.xxxx		x.xxxx
	LS Mean Estimate (SE)	x.x (x.xx)	x.x (x.xx)	x.x (x.xx)		x.x (x.xx)		x.x (x.xx)	
	95% CI	(x.x, x.x)	(x.x, x.x)	(x.x, x.x)		(x.x, x.x)		(x.x, x.x)	
Usability	Number of Assessments	xx	xx	xx	x.xxxx		x.xxxx		x.xxxx
	LS Mean Estimate (SE)	x.x (x.xx)	x.x (x.xx)	x.x (x.xx)		x.x (x.xx)		x.x (x.xx)	
	95% CI	(x.x, x.x)	(x.x, x.x)	(x.x, x.x)		(x.x, x.x)		(x.x, x.x)	
CUSQ Composite Scores [2]									

Variable Name	Variable Short Description	Length	Format	Definition	Origin
MATCHID	Paired propensity score matching ID. Two observations will be assigned to each unique value of MATCHID using the procedure employed to perform the propensity score match.	8	8.	= MATCHID from the output dataset produced by the propensity score matching procedure. If a complete match is obtained producing 164 subjects, the dataset will have 82 unique values of MATCHID with 2 records per unique value of MATCHID.	Derived
PS	Propensity score	8	8.	Propensity score generated by matching procedure.	Derived

Power Calculations

- BCRT offers support for simple and complex power analysis
  - Traditional statistics with defined solutions (e.g., t-test, ANOVA, general linear models)
  - Methods that require advanced tools (e.g., stepped wedge, multiple-period cross-over)
  - Techniques that require bespoke simulations (e.g., latent variable models and generalized linear mixed models)

- Power analyses are easiest to conduct with pilot data that closely mirrors the future study
  - Also possible with estimates of reasonable sample and effect size

- Some examples of recent BCRT support:
  - Stepped Wedge Trial (e.g., iCare)
    - https://clusterrcts.shinyapps.io/rshinyapp/
  - GLMM (e.g., MyPeeps Mobile LITE)
    - simr R Package
    - MPlus

Survey Methodology

- The BCRT has extensive experience with survey construction to make sure you are getting accurate responses from your participants
  - We also can provide early or mid-point collection assessments of your survey tool

- Most experienced with REDCap for data collection
  - Including Mobile REDCap

Geospatial Sampling

- Using geospatial attributes for sampling
  - Examples: inclusion of specific populations, or spatial proximity
  - Utilizing secondary sources, such as the American Community Survey (ACS)

Research Example:

Screening for Coronavirus Antibodies in Neighborhoods (SCAN)

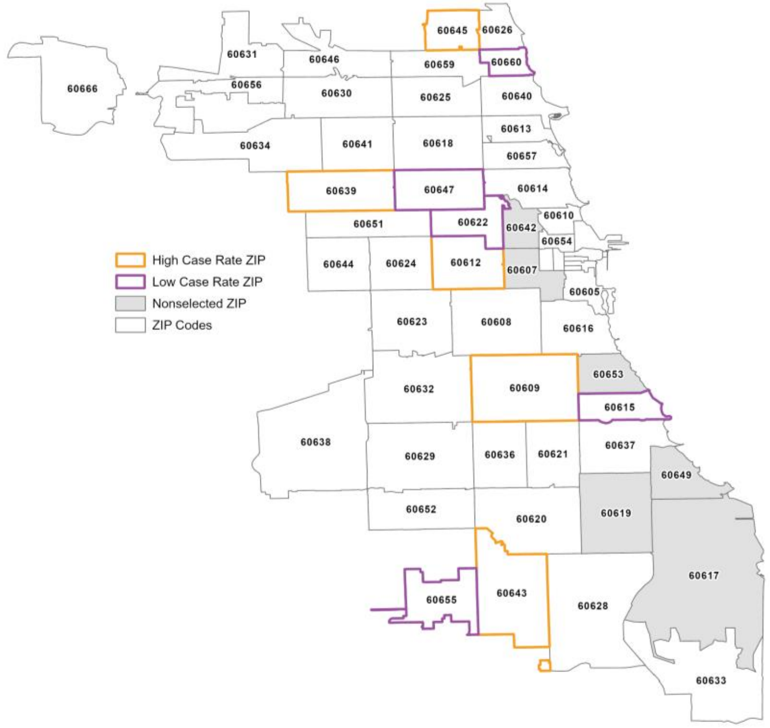
Geospatial Sampling: SCAN

SCAN was trying to find out how many people in specific areas of Chicago were exposed to COVID-19 by testing for antibodies in 2020-21.

- Interested in sampling adjacent ZIP codes in Chicago with differing COVID-19 case rates as reported by CDPH
- Pulled CDPH COVID-19 ZIP code data, Chiago Health Atlas, and American Community Survey (ACS) data to help determine which ZIP code pairs to use

Geospatial Sampling: SCAN

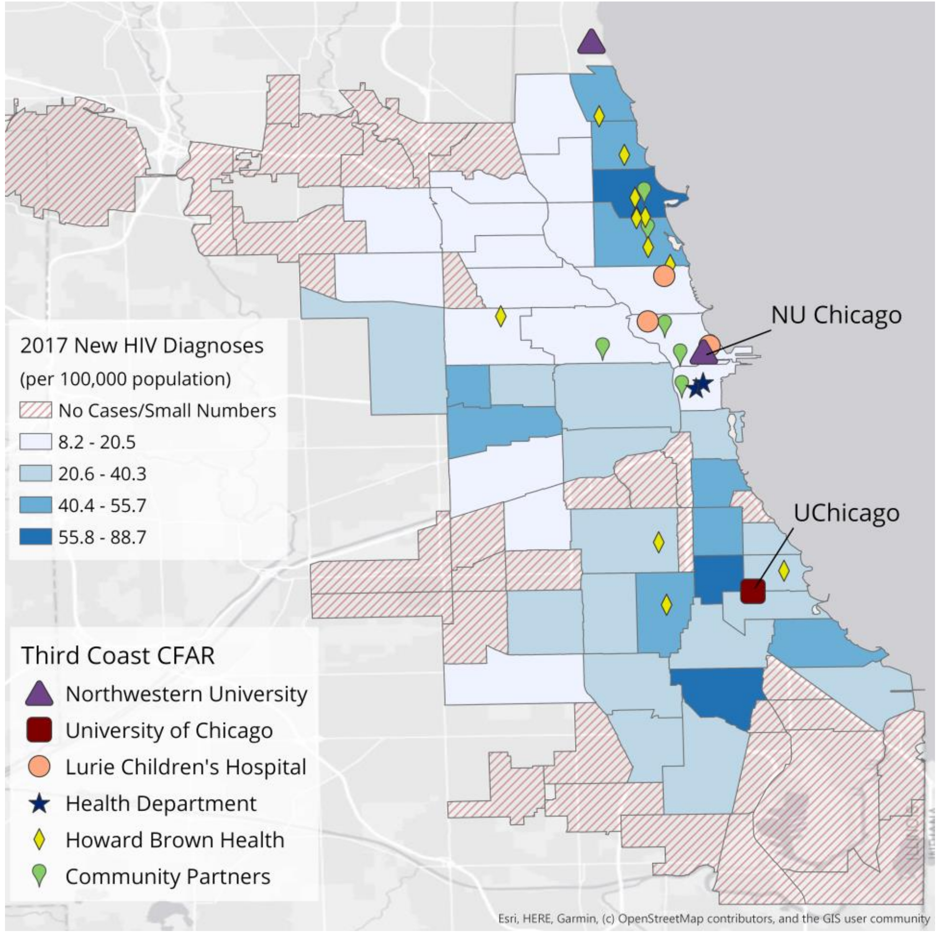
- Seven options for ZIP code pairs, some with more than two ZIPs to consider
  - Reviewed reported case rates from CDPH
  - Reviewed ACS measures like %workers using public transit to work, %occupied housing with 1.51 or more occupants per room, %uninsured, %below poverty
  - Reviewed health measures from Chicago Health Atlas (CDPH)
- Selected 5 ZIP code pairs, 10 ZIPs in total



Maps for Grant Applications

Including maps showing:

- Sampling locations
- % or rates of specific populations
- Location of partner orgs or schools



Post-Award Services

- Survey Programming & Data Management
- Quantitative Methods
  - Traditional
  - Spatial
  - Networks

Traditional Statistics

- Multiple Imputation
- Hierarchical Modeling

Multiple Imputation: Overview

- Iterative Stochastic Imputation

Multiple Imputation: Overview

- Iterative Stochastic Imputation
- Regression/Conditional Mean Imputation
- Outcome = $\alpha + \beta_0$(Non-Missing Variable) + β_1(Missing Variable)
- Approach replaces missing values with predicted values.
- Strength: Uses an established model-defined relationship elaborated to impute missing values.
- Problem: Artificially reduces estimates of variability for parameter estimates.

Multiple Imputation: Overview

- Iterative Stochastic Imputation
- Regression/Conditional Mean Imputation
- Outcome = α + β_0(Non-Missing Variable) + β_1(Missing Variable)
- Approach replaces missing values with predicted values.
- Randomly draws a residual/error term assuming N(0,1).
- Helps but estimates of variability will still be attenuated.

Multiple Imputation: Overview

- <u>Iterative Stochastic Imputation</u>
- Multiple imputation applies stochastic imputation a number of times.
- Distribution of non-missing data is used to estimate multiple values that when taken together reflect the uncertainty around the true value.

Multiple Imputation: Overview

- Imputation Phase: Missing data filled in with estimated data. Repeat this X times. Use all variables that need to be imputed for all analyses producing X datasets.

- Analysis Phase: Model of interest for the analysis is fit for each of the X datasets.

- Pooling Phase: Parameter Estimates (Beta Coefficients) and associated standard error estimates from each of the X iterations of the analysis phase are used to develop pooled estimates of effect and the variability of effect.
  - One guideline: X = % of observations for which complete data are not available

Multiple Imputation: Practice

- Keep It Up! (KIU!) Study
  - Web-based multimedia HIV prevention intervention for YMSM 18-29 y/o
  - Implementation: Community-Based Organization (CBO) delivered vs. a Direct-to-Consumer (DTC) model
  - Obtain estimates of HIV infections averted for each arm and compare cost-effectiveness
- COVID: led to large degree of missingness in survey and laboratory data
- Analysis Ongoing

have fun. stay safe.

keep it up!

Hierarchical Modeling: Overview

- Regression technique designed to deal with data that are inherently clustered.
  - Analytic units naturally nested within other units of interest.
    - Patients treated by a physician.
    - Patients seen at a hospital.
    - Hospitals within a state.
    - Students within schools.
- Acknowledges correlation of data that exists within clusters.
  - Data within a given cluster behave similarly
  - Behave similarly in a way that data from separate clusters do not.
  - Cluster membership has an important effect on your outcome of interest

Hierarchical Modeling: Simple Linear Regression

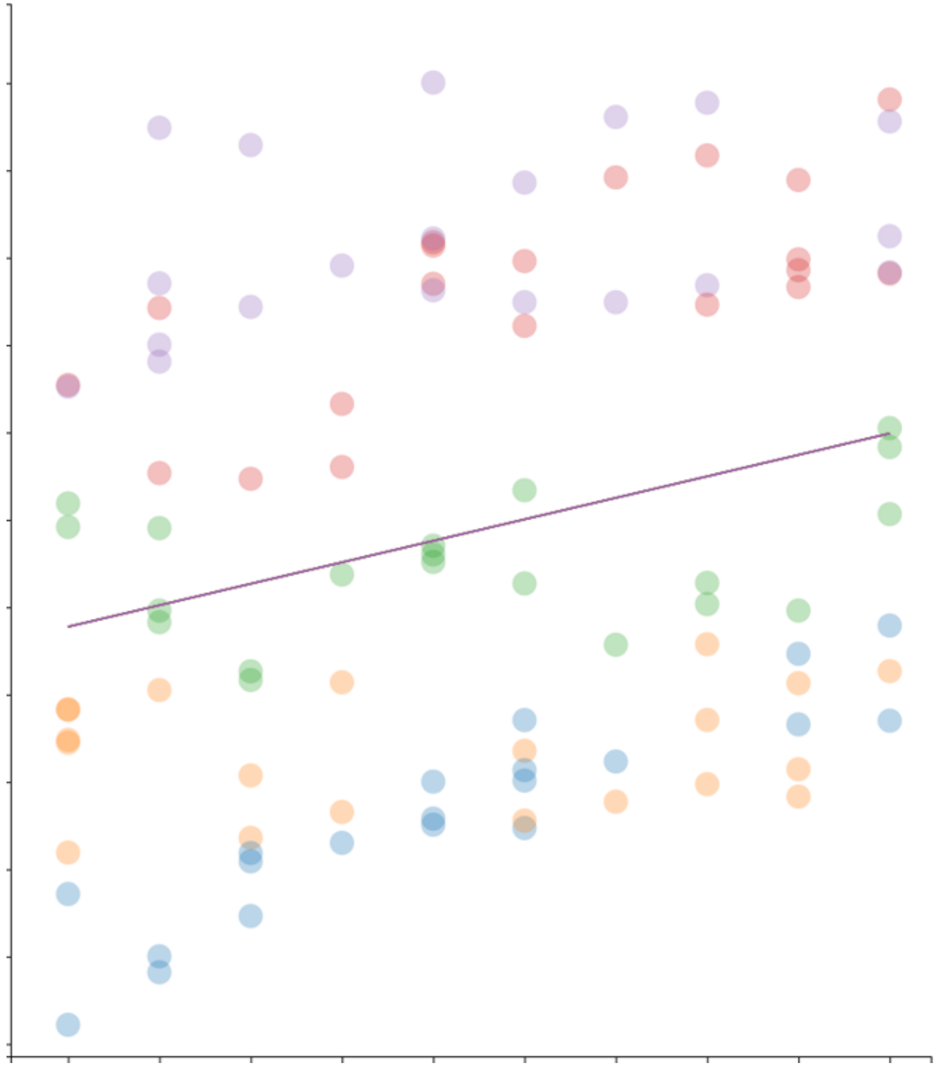
- <u>Fixed-Effects Model</u>
- Relationship is "fixed"

  Each point is represented by the same straight line.

- Model assumes "independence"

  Each point is independent or uncorrelated with every other point in the graph

- Violations of this assumption can lead to spurious results.

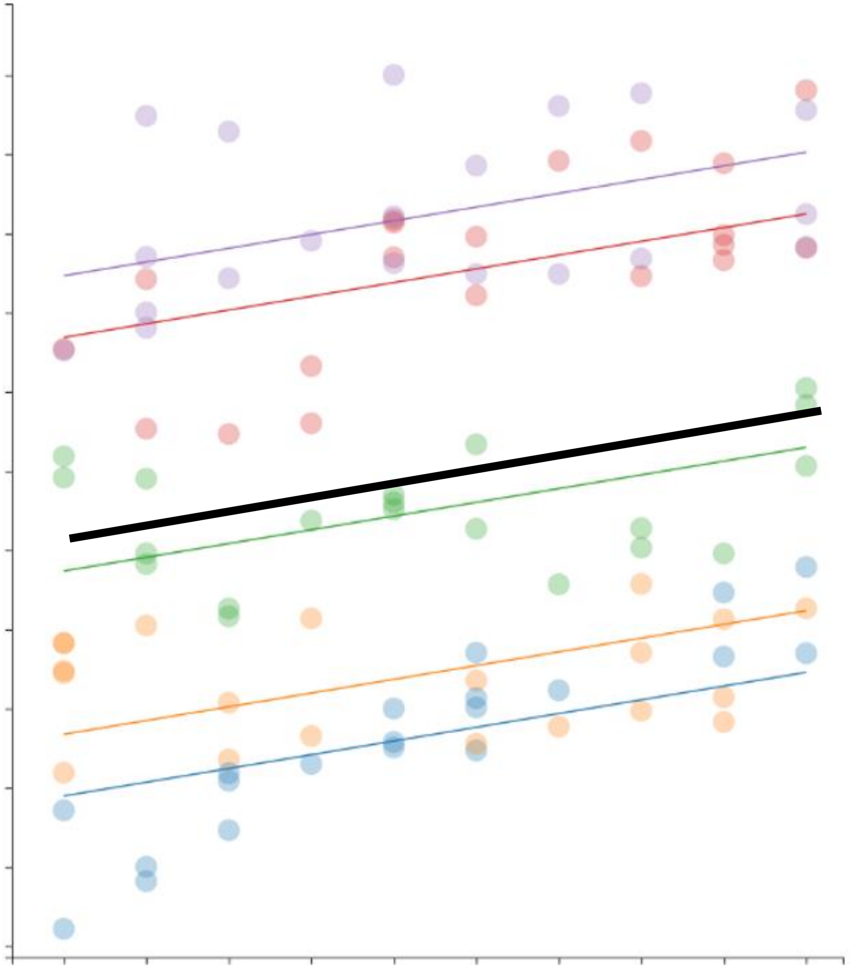


http://mfviz.com/hierarchical-models/
2017

Hierarchical Modeling: Mixed Modeling

- <u>Random-Effects or “Mixed” Model</u>
- Characterize the difference the cluster-specific regression lines and the population regression line.
- Quantify the relative magnitude of our outcome by cluster.

- Estimates are adjusted for the effects of all of the independent variables in our model.
- Fixed effects can characterize attributes as the patient level (Level 1) or at the cluster level (Level 2)
- Hierarchy can include more than 2 nested levels: (patient -> hospital -> state)

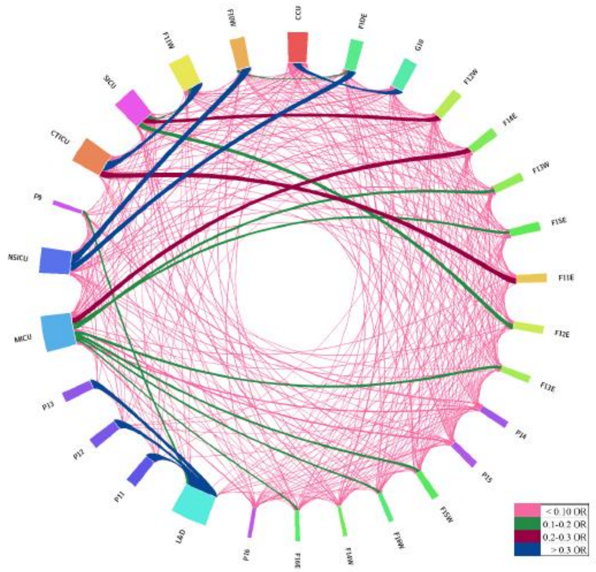


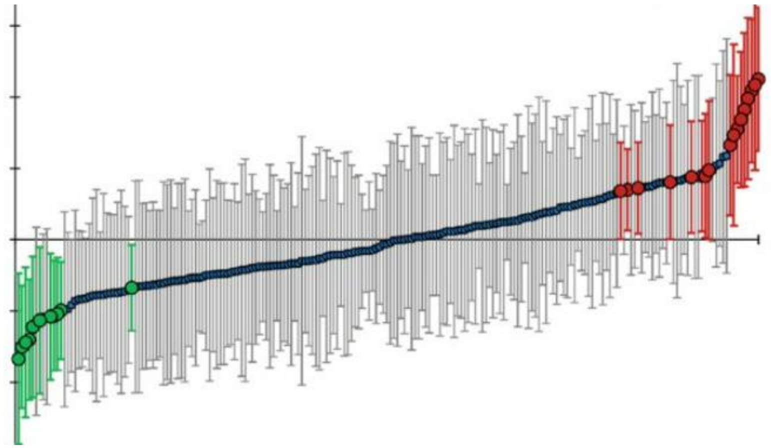
http://mfviz.com/hierarchical-models/
2017

Hierarchical Modeling Example: Intrahospital Transfer and Opportunistic Infection

- <u>Outcome</u>: Occurrence of opportunistic infection (VRE, CRAB, CRE, MRSA, *C. diff*)
- Clusters defined using a 3 level-hierarchy: encounter within patient within inpatient-unit-pair
- <u>Encounter level fixed effects</u>
  Transferred to another in-patient unit (Yes/No)
  Length of stay
- <u>Patient-level fixed effects</u>
  Age
  Gender
  Race
- <u>Unit-Pair level fixed effect</u>
  Overlap ratio =
  (2 x encounters shared b/t unit 1 and unit 2)/
  (total encounters unit 1 + total encounters unit 2)
- <u>Random intercept</u>

  Exponentiated random intercepts can be used to identify outlier unit-pairs





Kelly et al. In progress.

Spatial Statistics

Statistical tools to analyze patterns, distributions, relationships. Unlike traditional methods, they incorporate space directly into the equations.

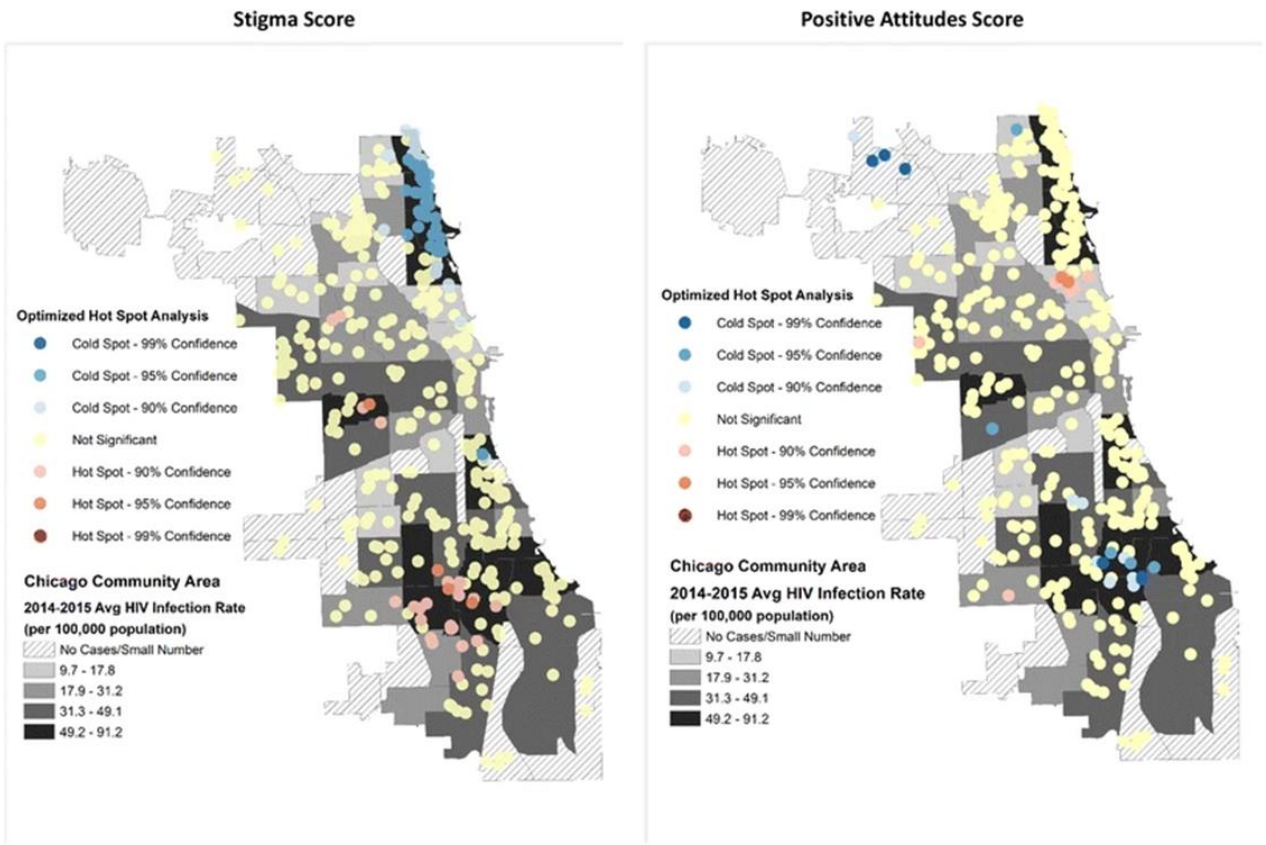
- Analyzing Patterns
    - Testing for random chance of clustering
        - Spatial Autocorrelation (Polygons)
        - Average Nearest Neighbor (Points)
- Mapping Clusters
    - Optimized Hot Spot Analysis - Identifying statistically significant hot and cold spots for a specific attribute
    - Emerging Hot Spot Analysis - Identifying trends in statistically significant hot and cold spots for a specific attribute over time (using space-time cubes).

Spatial Statistics: Optimized Hot Spot Analysis

Research Example:

RADAR – PrEP

- Optimized Hot Spot Analysis looking at an index score (continuous) for PrEP Stigma or Positive Attitudes toward PrEP.



Mustanski, B., Ryan, D.T., Hayford, C. *et al.* Geographic and Individual Associations with PrEP Stigma: Results from the RADAR Cohort of Diverse Young Men Who have Sex with Men and Transgender Women. *AIDS Behav* 22, 3044–3056 (2018). https://doi.org/10.1007/s10461-018-2159-5

CDPH Collaboration

CFAR has a research relationship with CDPH

- Honest Broker model
  - Research can be done at CDPH by honest broker, with CDPH approval
  - Data requests for grant applications, or model inputs

Project Examples:

- Next-Generation Phylodynamics-Targeted Partner Service Models for Combined HIV Prevention (P2M)
- CDPH Demonstration Project – CDC funded
- Illinois/Chicago Getting to Zero Dashboard

CDPH Collaboration

CFAR has also contributed to CDPH efforts

- GIS portions of CDPH HIV + STI Annual Reports
- Assisting Epidemiology team with data tasks
- Connecting HIV/STI team with Northwestern and UChicago researchers and clinicians
- GIS Training for HIV/STI Epidemiology team

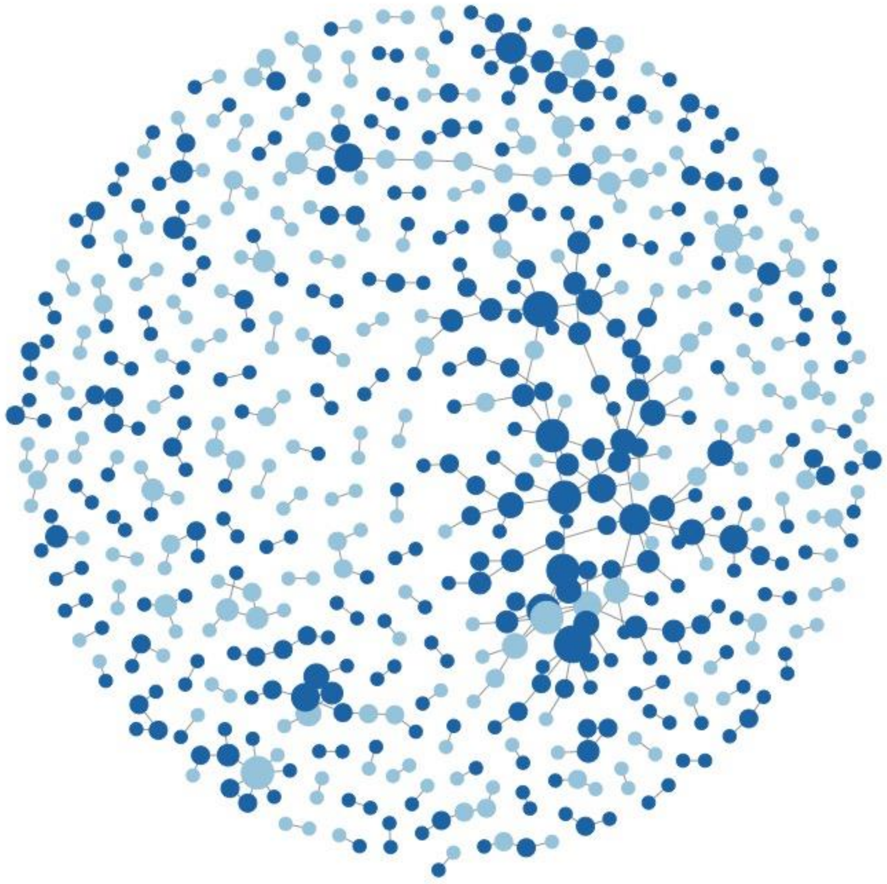
Networks

- Support for all stages and types of network analysis

- Network Measurement
  - Using traditional survey tools (e.g., REDCap, Qualtrics)
  - Using specialized tools (e.g., NetworkCanvas.com, EgoWeb)

- Network Analysis
  - Whole networks
  - Egocentric networks
  - Visualizations
  - Simple and Advanced analysis (e.g., exponential random graph models; ABM)

- CFAR Supported projects with network analysis
  - P2M (PI: Benbow, D'Aquila, Schneider)
  - RADAR (PI: Mustanski)
  - ChiSTIG (PI: Birkett)
  - PUG 2 (PI: Janulis)



Additional Services

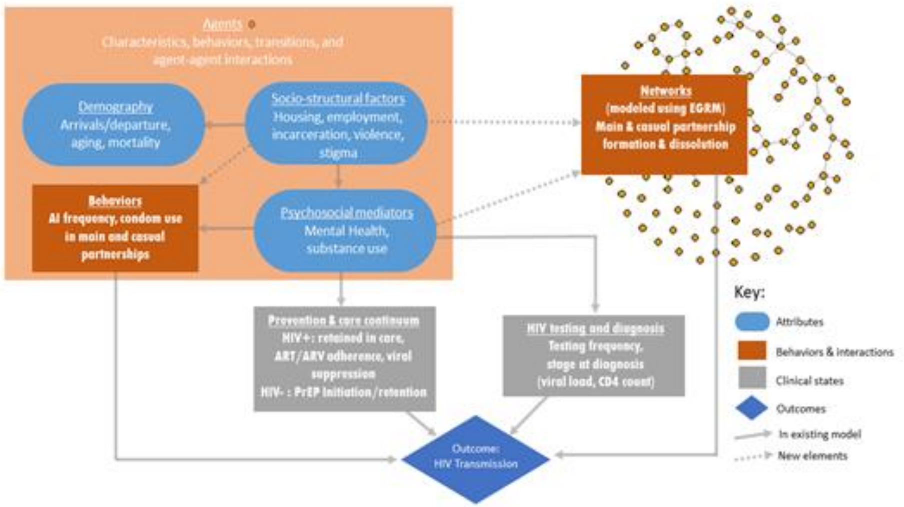
- Methods in Epidemiology and Biostatistics
  - Epidemiologic study design, sampling, survey design
  - Causal inference methods
  - Statistical methods for clustered and longitudinal data
  - Meta-analysis

- Study Designs
  - Sequential, Multiple Assignment, Randomized Trials (SMART)
    - SMART trial to address alcohol use among Black sexual minority men (PIs McNulty, Knox), submitted January 2023
  - Hybrid Effectiveness-Implementation trials
    - Recently funded R01MH131476 (PIs Schneider, Bouris): Implementation of a triadic network case management intervention for younger Black sexual minority men

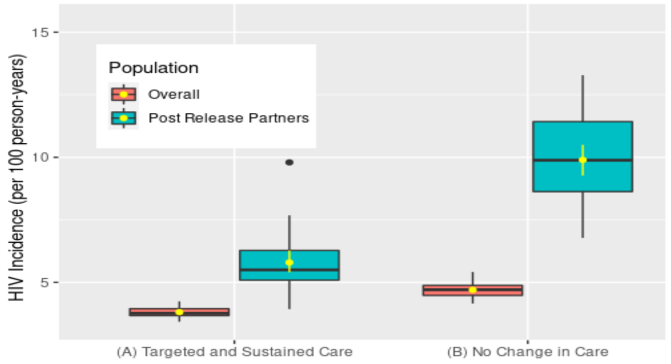
Agent-based, network, and complex systems modeling

- Agent Based Models (ABMs)
  - Computer simulation approach to modeling the dynamics of complex systems
  - Models represent social systems composed of agents that interact with and influence each other
  - Observe system level consequences of agent behaviors and interactions
  - Effects of interventions can be simulated under various assumptions in a virtual environment
- CFAR supported projects
  - P2M (PIs Benbow, Benbow, Schneider)
  - Modeling social determinants of health and HIV transmission (R21MH128116, PI Hotton)

Components of an Agent-Based Model



10-year HIV incidence rates under (A) targeted and sustained ART interventions for jail detainees and (B) standard care



Requesting CFAR Services

- https://www.thirdcoastcfar.org/services/
  - Service Request Form: https://app.smartsheet.com/b/form/daae3fb20a3042888170cf2b7c8b3555

  - Email Angel Avina: angel.avina@northwestern.edu

Contacts

- Pat Janulis: patrick.janulis@northwestern.edu
- Anna Hotton: ahotton@medicine.bsd.uchicago.edu
- Christina Hayford: christina.hayford@northwestern.edu
- Susheel Reddy: susheel.reddy@northwestern.edu
- Angel Avina: angel.avina@northwestern.edu