Design Considerations for Running Health System Based Trials Through the Electronic Health Record

Benjamin A. Goldstein, PhD, MPH

ben.goldstein@duke.edu

Department of Biostatistics & Bioinformatics
School of Medicine
Duke University

April 17th, 2019

Why We Want to use EHRs for Clinical Research

- Data readily available
- Often 100,000's of Patients
- Information collected over a variety of fields
- Can study just about any clinical outcome
- Representative Population

Why We May *Not* Want to use EHRs for Clinical Research

Data are not collected for research

- Data exist in disparate places
- All patients have different pieces of information
- Observational Data

Four Ways EHR Data Differ from Traditional Clinical Data

1. We don't have everything we want
2. Outcomes are not defined - need to phenotype data
3. Data are both longitudinal and cross-sectional
4. Data not observed randomly - Informed Presence

Challenge 1:
We don't have everything we want

- Patients may seek care at multiple facilities
- Most social health information is not recorded or reliable
- Cannot expect death is reliably captured
  - Most people don't die in the hospital
  - Preliminary work suggests EHRs have only 20% sensitivity

Addressing Incompleteness Via Design

- Define local patient population
  - Live in the catchment of the health system
  - Require a certain a number of primary care appointments before eligible for study
- Contextual and proxy information can be linked in
  - Neighborhood for SES
  - Claims data for additional encounters
  - NDI/SSDI for death

Challenge 2

Issues of Data Definition: What is a Diabetic?

Research and applications

A comparison of phenotype definitions for diabetes mellitus

Rachel L Richesson,[1] Shelley A Rusincovitch,[2] Douglas Wixted,[3] Bryan C Batch,[4] Mark N Feinglos,[4] Marie Lynn Miranda,[5] W Ed Hammond,[2,6] Robert M Califf,[3,7] Susan E Spratt[4]

▸ Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/amiajnl-2013-001952).

[1]Duke University School of Nursing, Durham, North Carolina, USA
[2]Applied Informatics Research, Duke Health Technology Solutions, Durham, North Carolina, USA
[3]Duke Translational Medicine Institute, Durham, North Carolina, USA
[4]Division of Endocrinology, Metabolism and Nutrition, Department of Medicine, Duke University School of Medicine, Durham, North Carolina, USA
[5]Department of Pediatrics, School of Natural Resources and Environment, University of Michigan, Ann Arbor, Michigan, USA
[6]Duke Center for Health Informatics, Durham,

ABSTRACT

Objective This study compares the yield and characteristics of diabetes cohorts identified using heterogeneous phenotype definitions.

Materials and methods Inclusion criteria from seven diabetes phenotype definitions were translated into query algorithms and applied to a population (n=173 503) of adult patients from Duke University Health System. The numbers of patients meeting criteria for each definition and component (diagnosis, diabetes-associated medications, and laboratory results) were compared.

Results Three phenotype definitions based heavily on ICD-9-CM codes identified 9–11% of the patient population. A broad definition for the Durham Diabetes Coalition included additional criteria and identified 13%. The electronic medical records and genomics, NYC A1c Registry, and diabetes-associated medications definitions, which have restricted or no ICD-9-CM criteria, identified the smallest proportions of patients (7%). The demographic characteristics for all seven phenotype definitions were similar (56–57% women, mean age

populations. Furthermore, standard
nitions can streamline the develop
registries from healthcare data, and
inclusion criteria to support regional
the identification of rare disease co
understanding of the populations
various phenotype definitions will
methods for identifying diabetes c
the rapid generation of patient regist
datasets with uniform sampling crit
comparative and aggregate analysis.
study presents and compares the size
tics of patient populations retrieved
phenotype definitions adopted from
betes registries and research networ
munity intervention program in
federal reporting standards.

BACKGROUND AND SIGNIFICANC
Diabetes diagnosis and manageme
Diabetes is a complex disease wit
types associated with different etiol

Table 1 Data domain criteria used in selected phenotype definitions

Phenotype definitions:	Data domain criteria: ICD-9-CM 250.xx	ICD-9-CM 250.x0 and 250.x2 (excludes type 1 specific codes)	Expanded ICD-9-CM Codes (249.xx, 357.2, 362.0x, 366.41)	HbA1c	Fasting glucose	Random glucose	Abnormal OGTT	Diabetes-associated medications*
ICD-9-CM 250.xx	●							
CMS CCW	▲✳\\		▲✳\\					
NYC A1c Registry				●				
Diabetes-associated medications								●
DDC		▲	▲	▲\\	▲\\	▲\\	▲\\	▲
SUPREME-DM	▲✳\\		▲✳\\	▲\\	▲\\	▲\\	▲	▲
eMERGE†		●✳\\		▲	▲	▲		▲

*Medications vary by phenotype definition and are listed for each in the supplementary appendix (available online only).
†The eMERGE phenotype definition consists of five case scenarios with varying combinations of criteria. Any instance of type 1 specific codes (ie, 250.x1, 250.x3) results in the exclusion of the patient.
●=Sole criteria.
▲=Optional criteria, one of many.
✳=Distinction made between inpatient and outpatient context.
\\ = Distinction made for multiple instances and/or time points.
CMS CCW, Centers for Medicare and Medicaid Services Chronic Condition Data Warehouse; DDC, Durham Diabetes Coalition; eMERGE, electronic medical records and genomics; HbA1c, hemoglobin A1c; ICD-9-CM, International Classification of Disease, revision 9, clinical modification; NYC, New York City; OGTT, oral glucose tolerance test; SUPREME-DM, Surveillance, Prevention, and Management of Diabetes Mellitus.

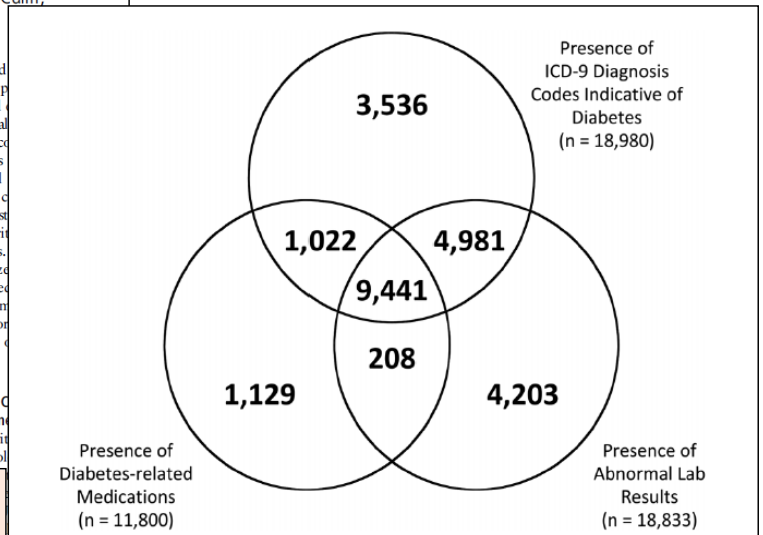


Figure 1 Overlap of diabetes cohorts identified from different categories of phenotype eligibility criteria; n=24 520 patients identified by criteria from any of the three categories.

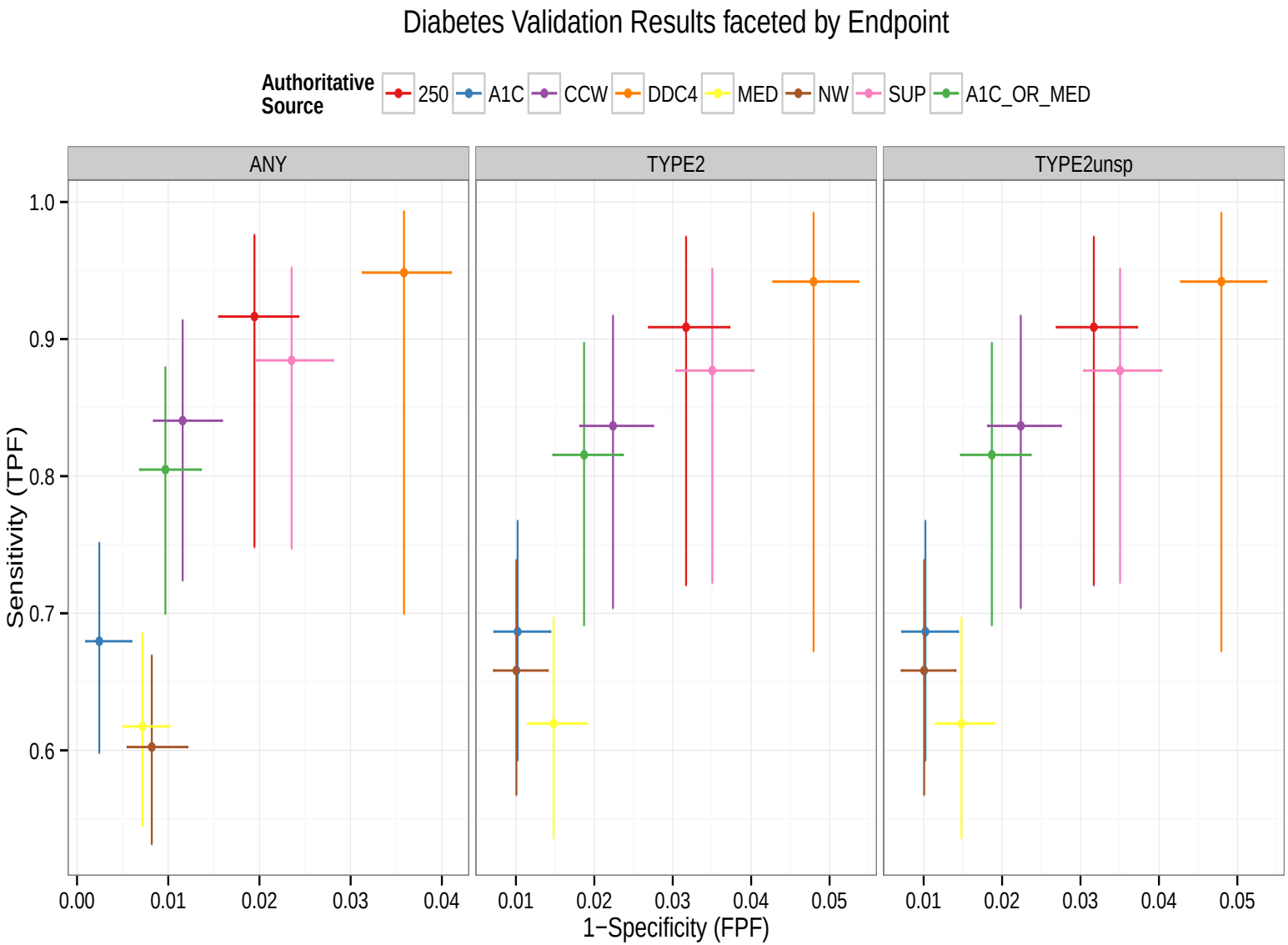
Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, Hammond WE, Califf RM, Spratt SE. A Comparison of Phenotyp Definitions for Diabetes Mellitus. J Am Med Inf Assoc 2013 (epub ahead of print). http://www.ncbi.nlm.nih.gov/pubmed/24026307

Issues of Data Definition: What is a Diabetic?

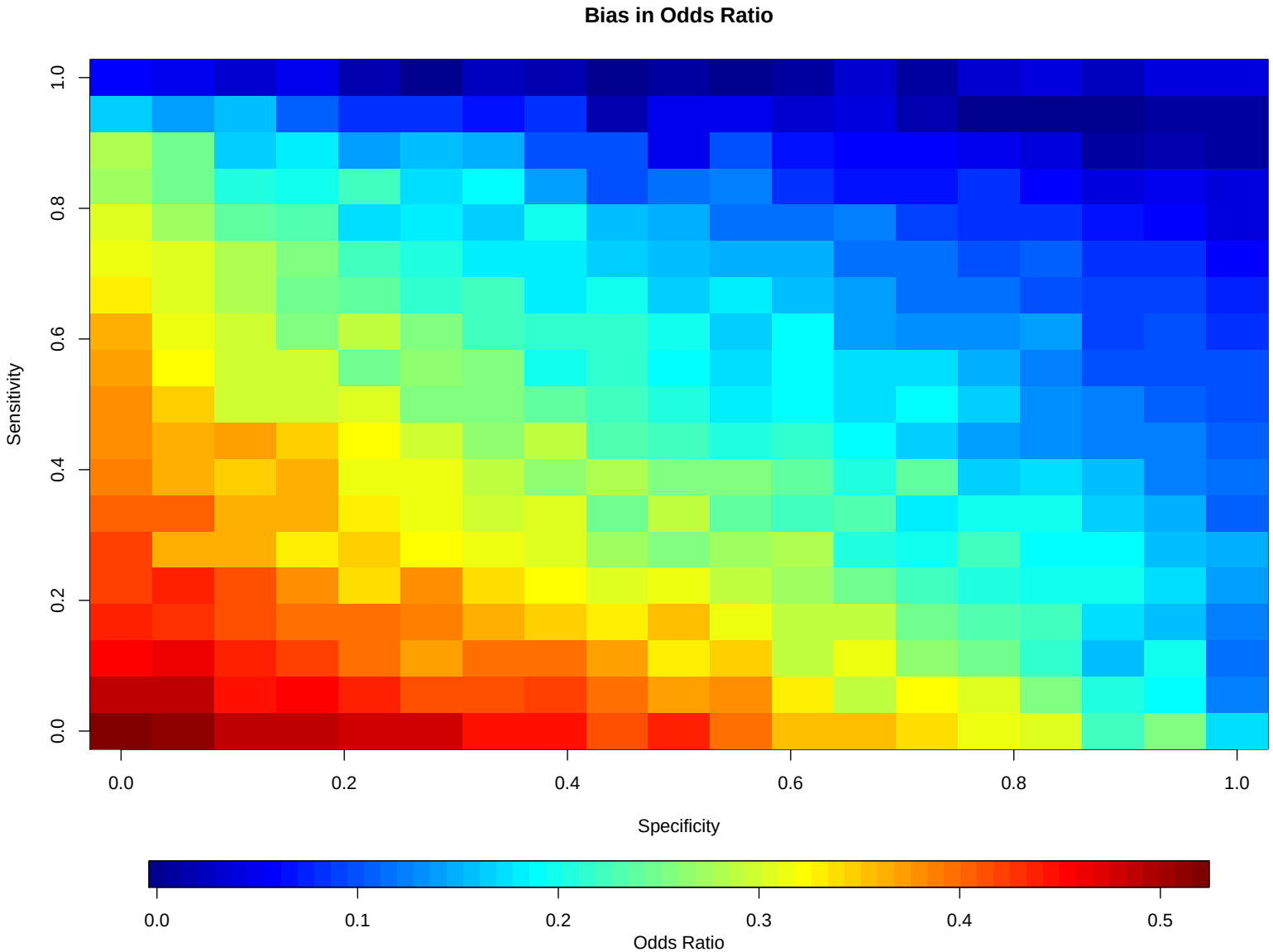
	ICD-9 250.xx	ICD-9 250.x0 & 250.x2 (exclude type I)	Expand. ICD-9 (249.xx, 357.2, 362.0x, 366.41)	HbA1c	Glucose	Abnormal OGTT	Diabetes Meds
ICD-9 250.xx	X						
CMS CCW	X*		X*				
NYC A1c Registry				X			
Meds							X
DDC		X	X	X	X	X	X
SUPREME-DM	X*		X*	X	X	X	X
eMERGE		X*		X	X		X

* Distinction between Inpatient and Outpatient Visits

Definition Differences



Impact of Poorer Definitions



CHALLENGE 3:
DATA ARE BOTH LONGITUDINAL AND CROSS-SECTIONAL

- EHR Data consist of *cross-section* of *longitudinal data*
  - Most data are stored in datamarts that cover fixed periods of time
- Need to use methods for longitudinal data to model updating exposures
  - We most often use time-varying Cox Models
  - Most analyses don't account for a patient's trajectory - just most recent value
- Since data are a cross-section no notion of time 0
  - Define "burn-in" periods to define eligibility
  - Use "burn-out" periods to define censoring

Challenge 4
Data are informatively observed: Informed Presence

- Collection of biases due to the fact that patients do not interact randomly with a health system
- Focus on what data are *observed* as opposed to what are *missing*

Three Types of Informed Presence

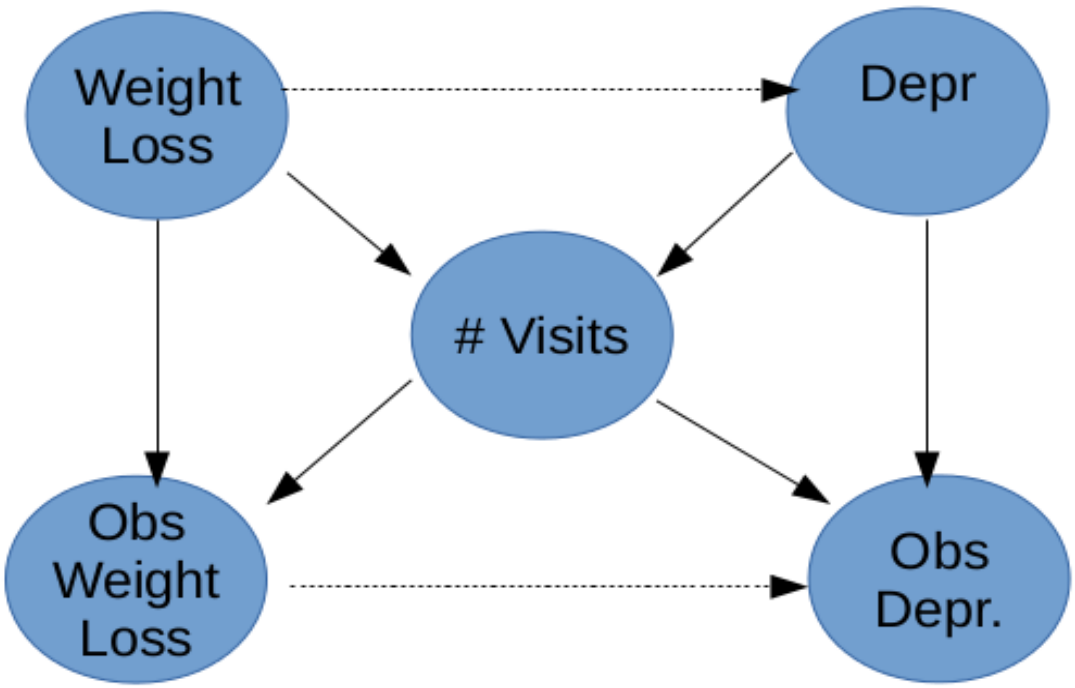
1. We know more about sicker patients
2. Where a patient seeks care is informative
3. Health status driving encounters

Informed Presence I:
Need to Account for Number of Encounters

Regression of Depression on Weight Loss

	Odds Ratio	Δ log(OR)	Δ OR
Minimally Adjusted	3.98 (3.81, 4.17)	—	—
+ No. Encounters	2.37 (2.26, 2.50)	-0.52	-1.61
+ Comorbidities	2.82 (2.69, 2.96)	-0.35	-1.16
+ No. Encounters & Comorb	2.30 (2.18, 2.42)	-0.55	-1.68

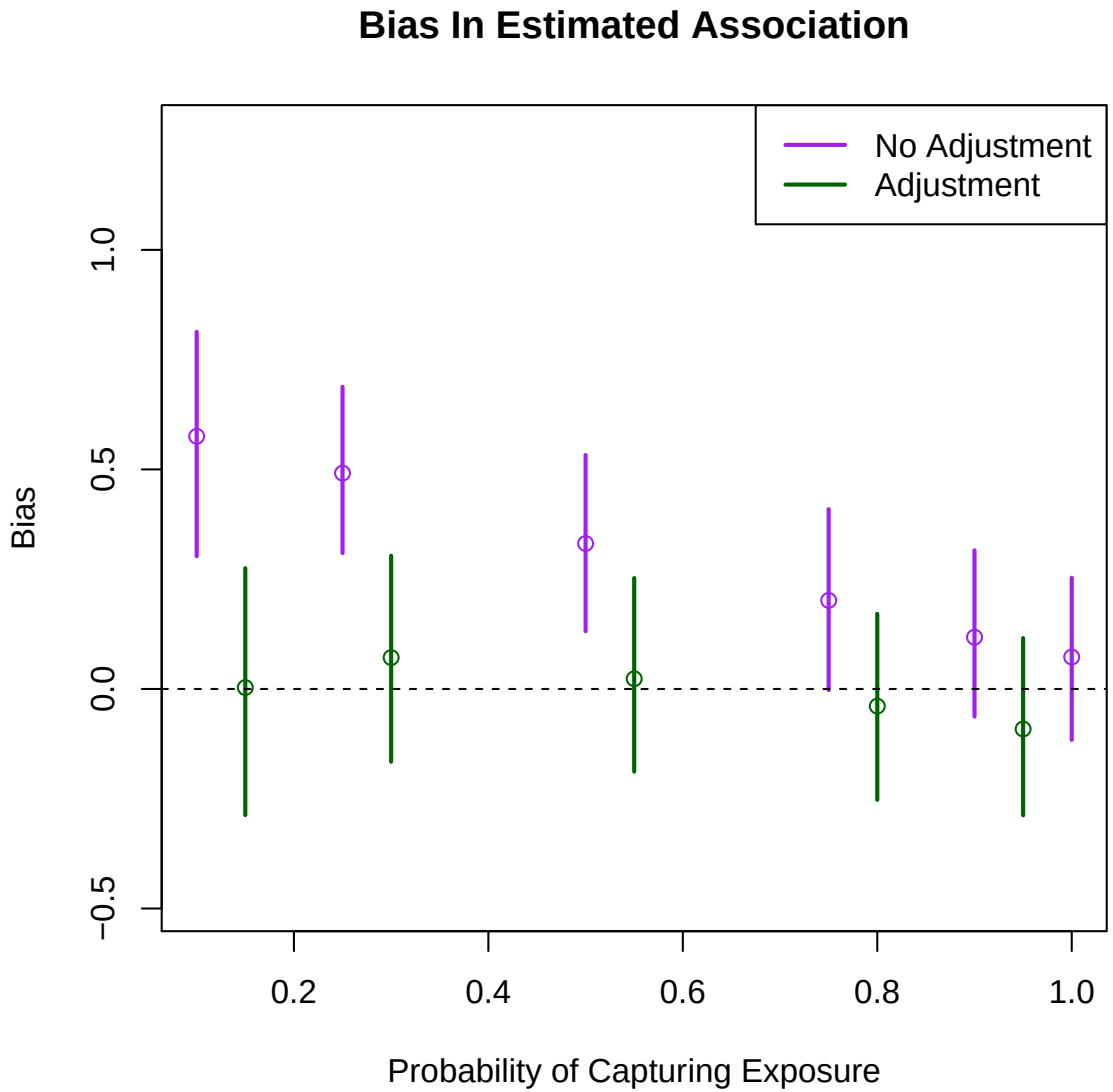
Number of Encounters Potential Confounder



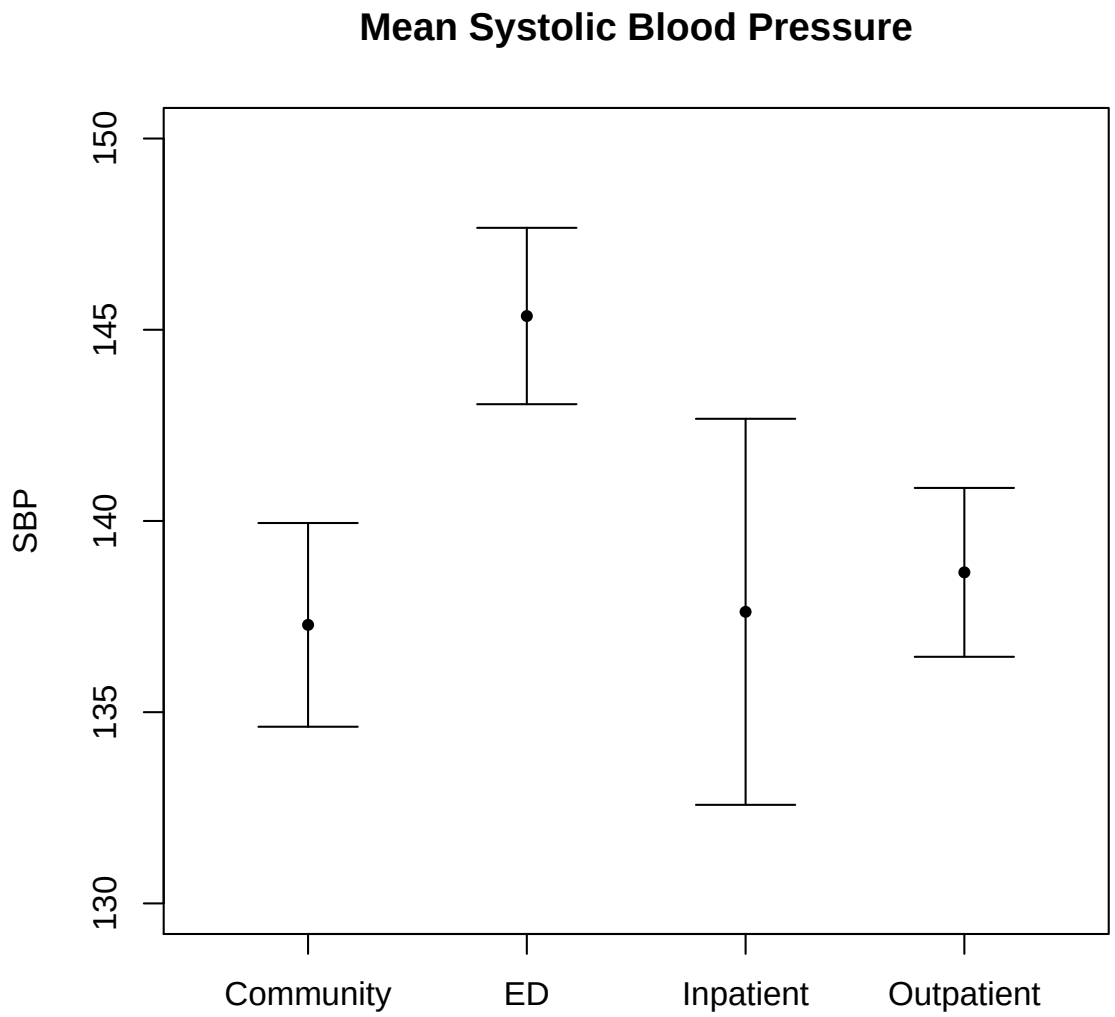
Need to Account for Number of Encounters

	Sensitivity	Median Number of Encounters Without Condition	With Condition
Depression	56.3%	6	38
Weight Loss	9.3%	7	45

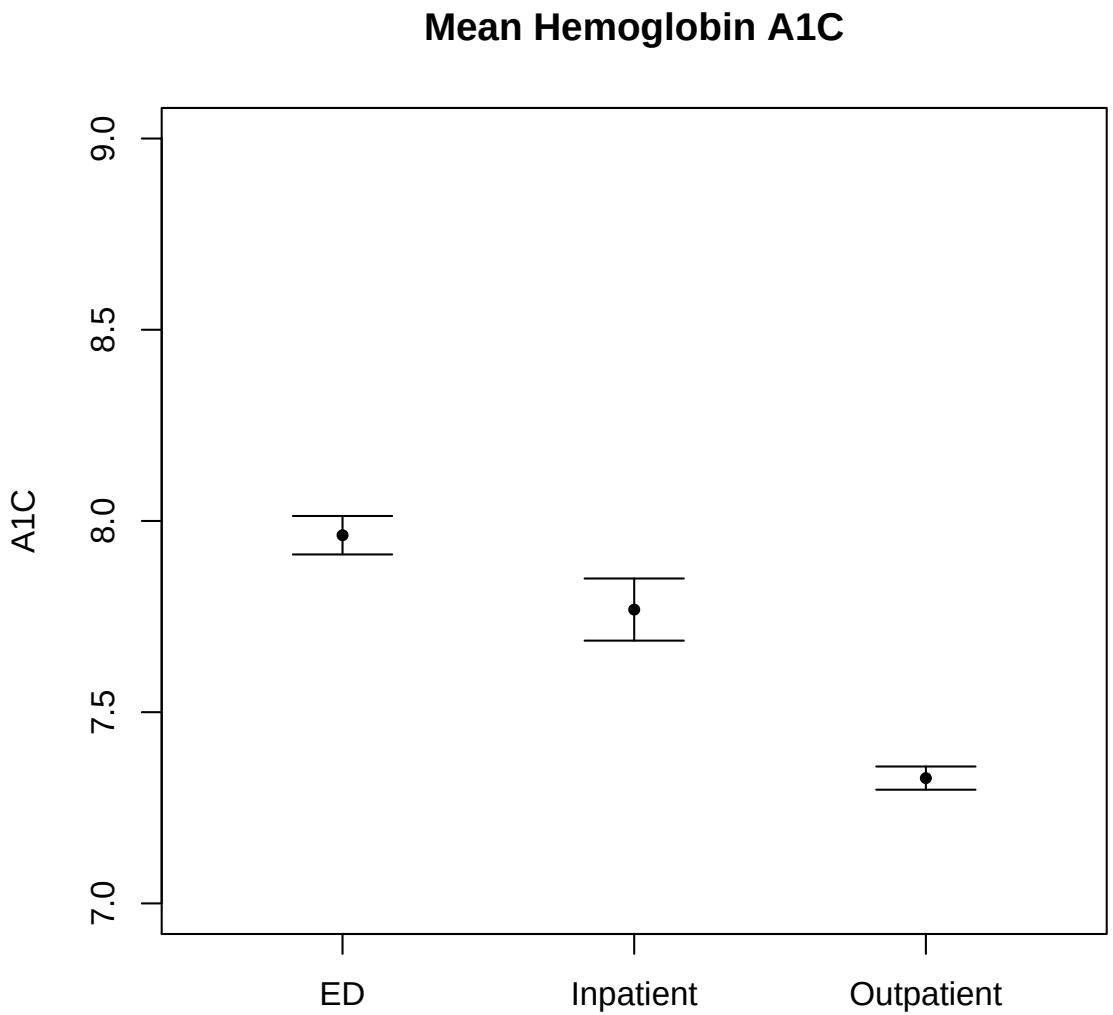
Number of Encounters Potential Confounder



Informed Presence II:
Where a Person Seeks Care is Informative



Where a Person Seeks Care is Informative



Location Impacts Inference

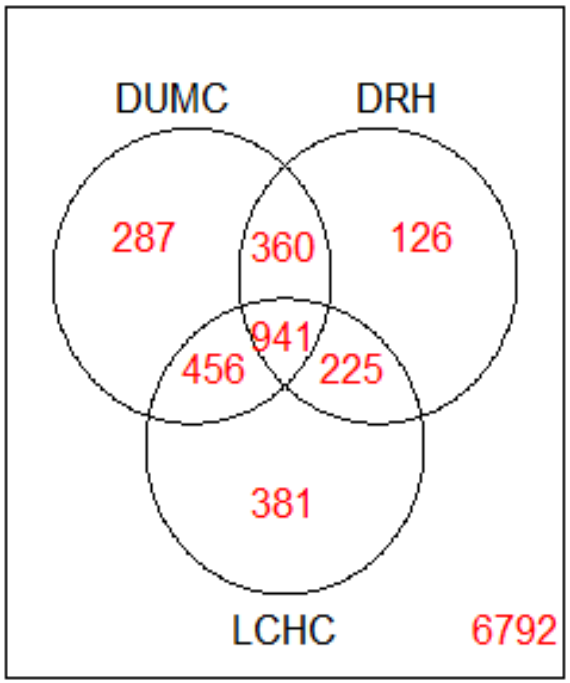
- Hazard Ratio for HgB A1C for time to Myocardial Infarction

Type	Hazard Ratio	P-value
Unadjusted	1.06 (1.01, 1.11)	0.026
Adjusted for Location	0.97 (0.92, 1.02)	0.178
OP Only	1.07 (1.00, 1.14)	0.044
ED Only	0.94 (0.89, 0.99)	0.022

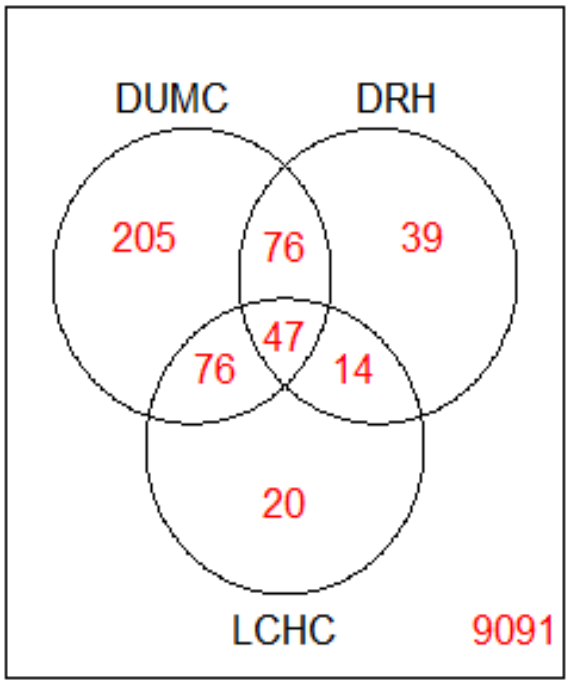
- Interaction between A1C and location

WHICH HOSPITAL A PATIENT USES IS INFORMATIVE

Diabetes
N=2,783



Cancer
N=477

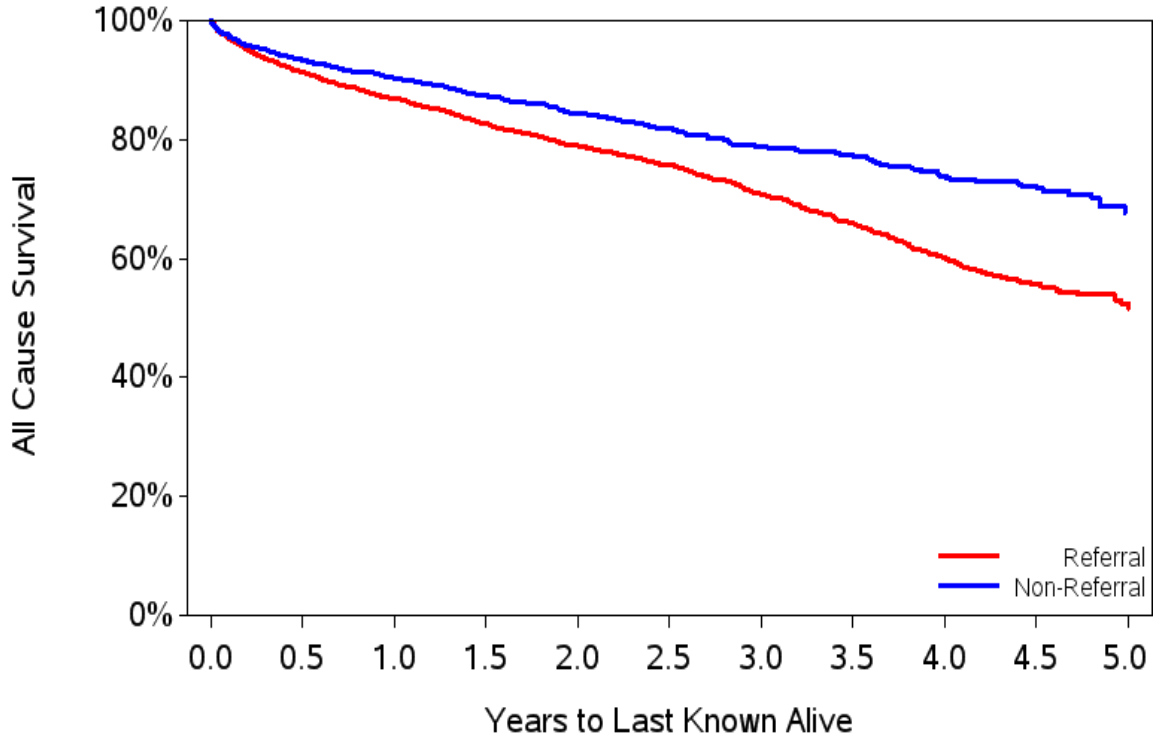


Facility Impacts Inference

- Odds Ratio for Cancer Status on Diabetes

Location	Odds Ratio	95% CI
All Facilities	1.69	(1.36, 2.10)
DUMC Only	1.46	(1.15, 1.87)
DRH Only	0.89	(0.63, 1.26)
LCHC Only	1.08	(0.74, 1.56)

Referral Hospitals are an *Admixed* Population



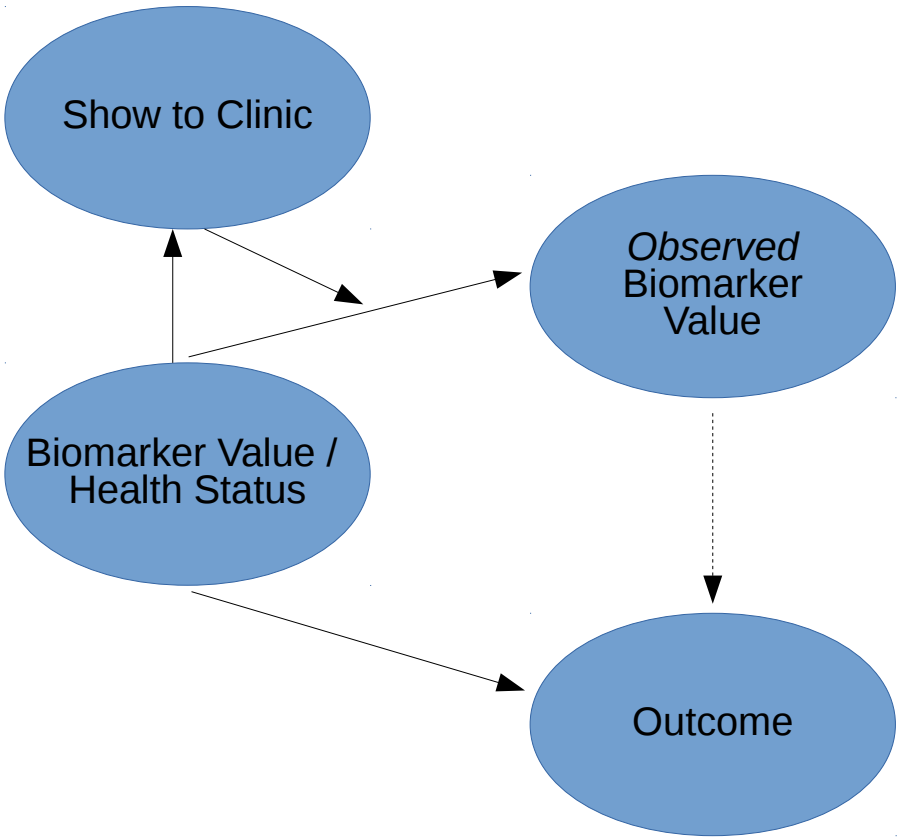
Number at risk											
Referral	5522	3307	2690	2159	1748	1360	995	697	474	282	65
Non-Referral	2114	1532	1318	1110	882	697	519	387	262	171	64

Admixture Bias

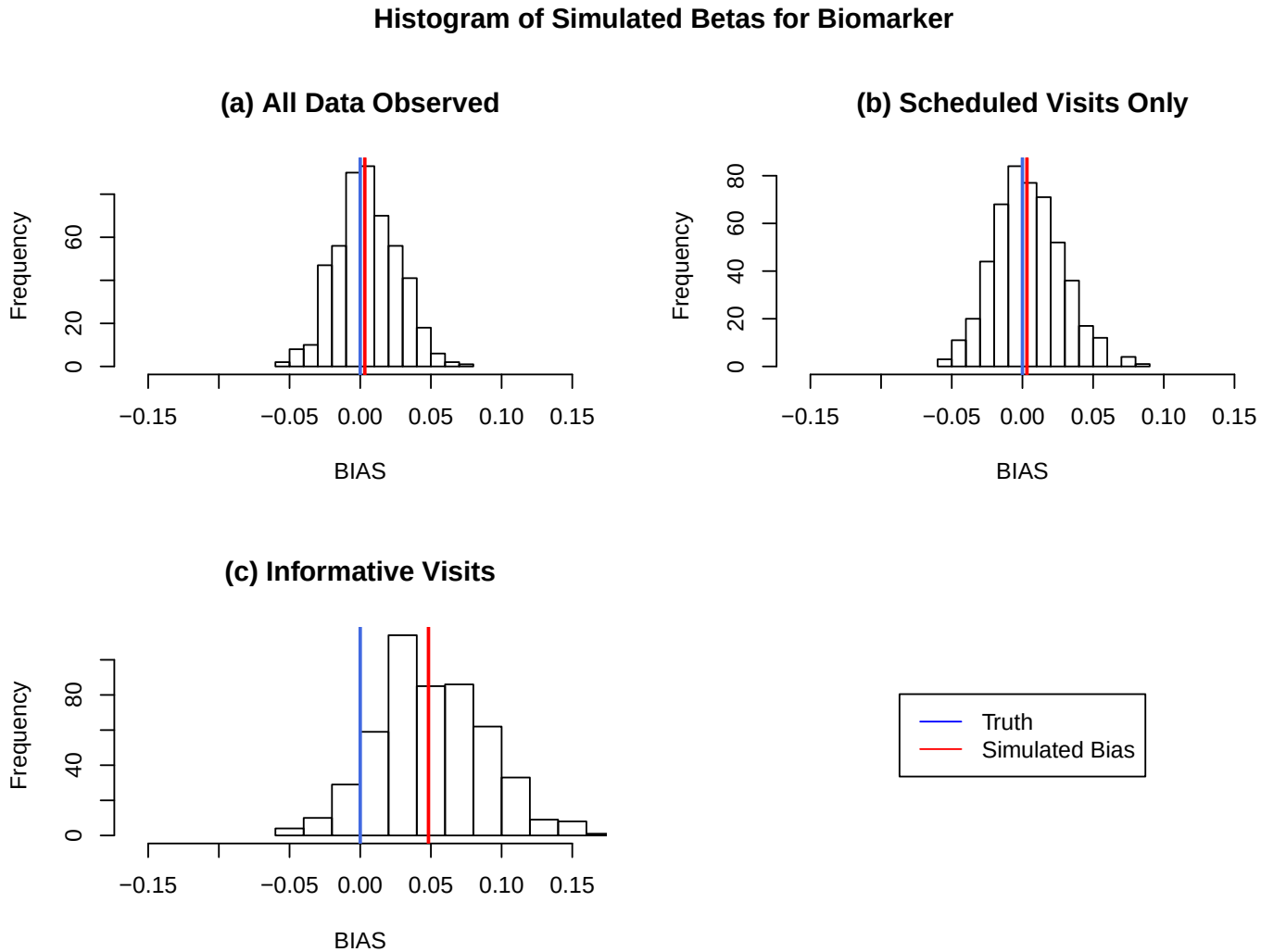
- Comparison of Local and Referral Patients at Cardiac Catheterization Lab

Local Patients	**Referral Patients**
Older	Younger
More Comorbidities	More severe valve disease
Disease due to ageing	Disease due systematic factors
Better outcomes	More follow-up procedures

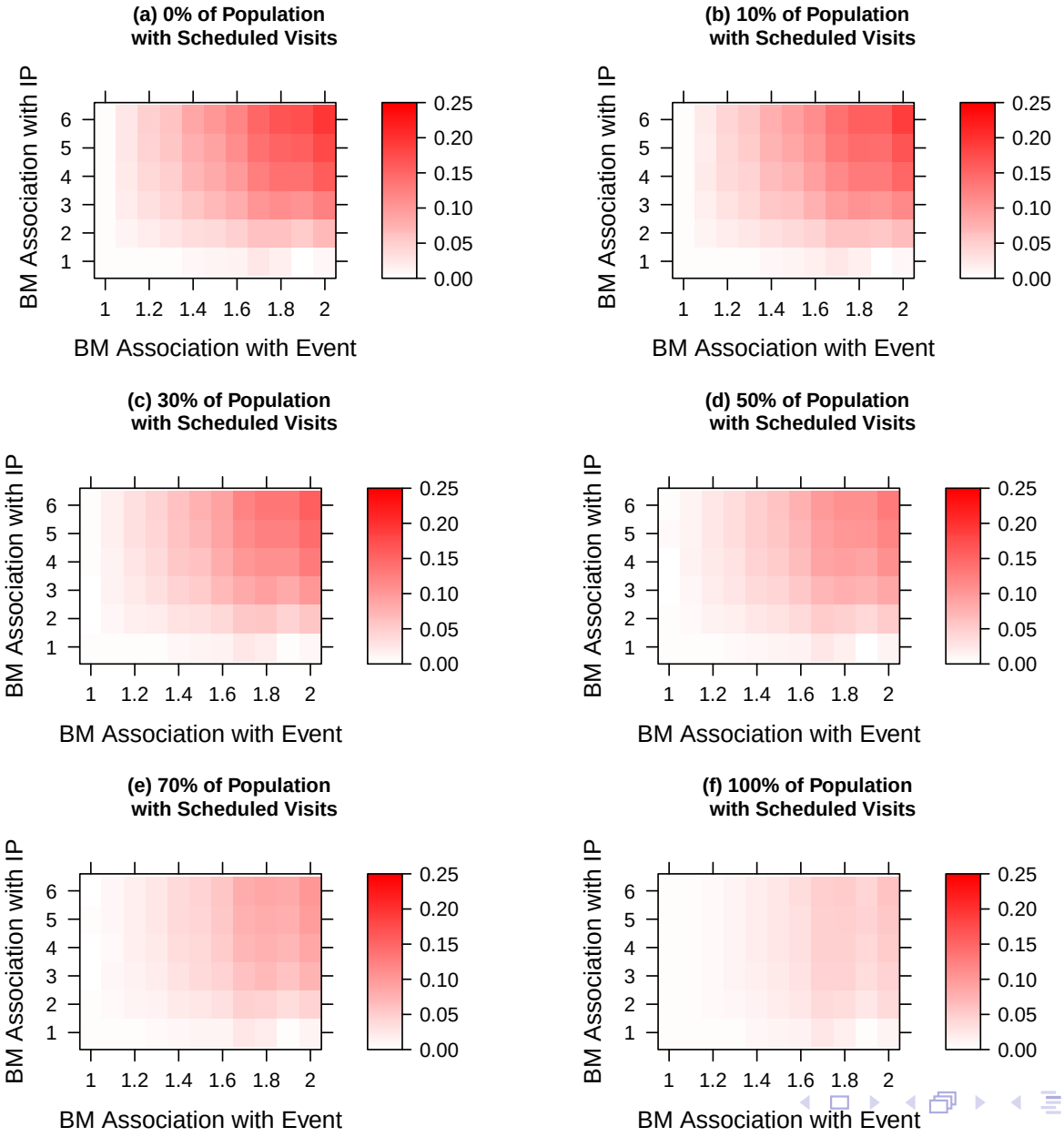
INFORMED PRESENCE III:
HEALTH STATUS DRIVING ENCOUNTERS



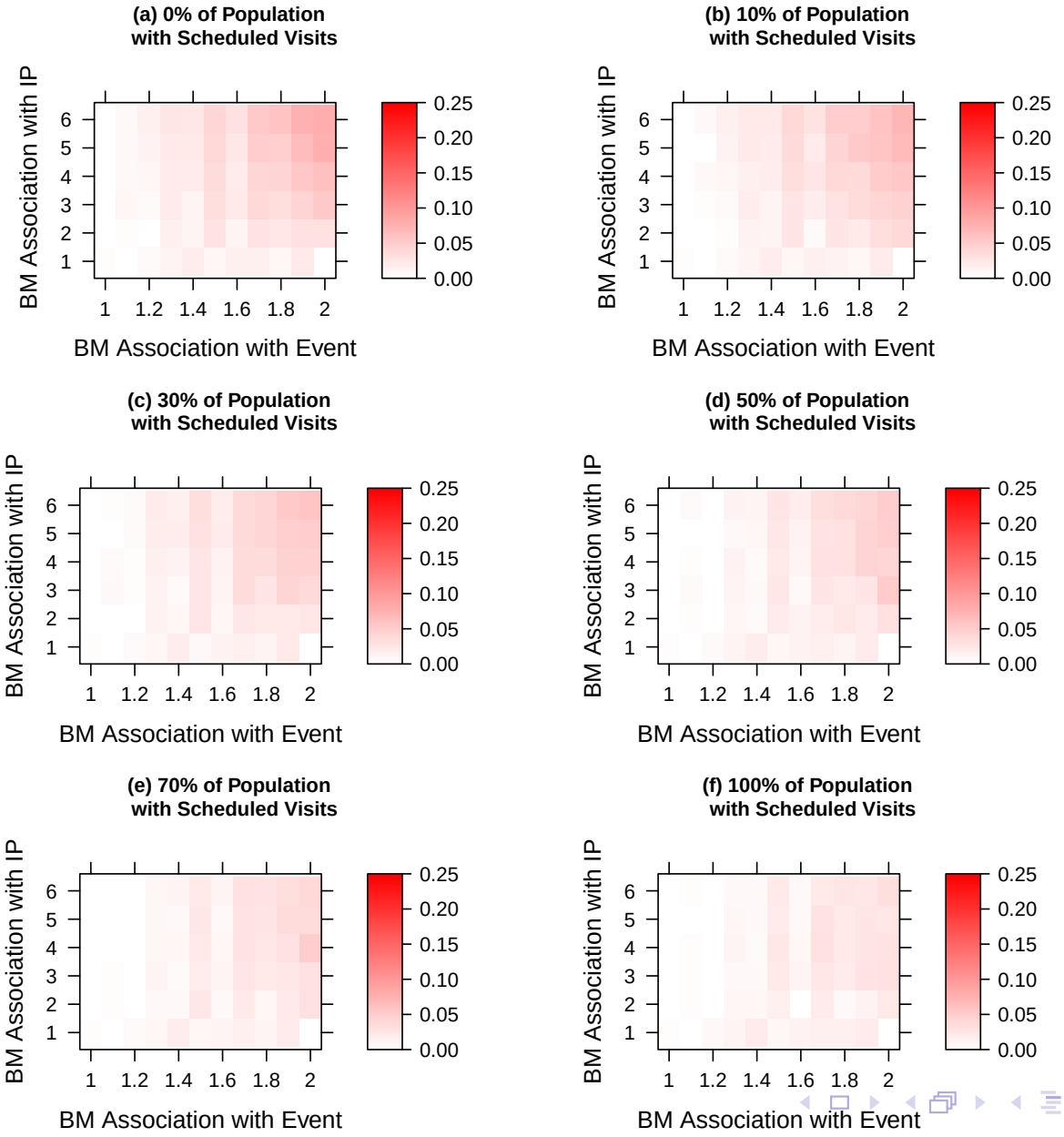
Impact of Informative Visit Process on Bias



Need an Underlying Association to Induce Bias



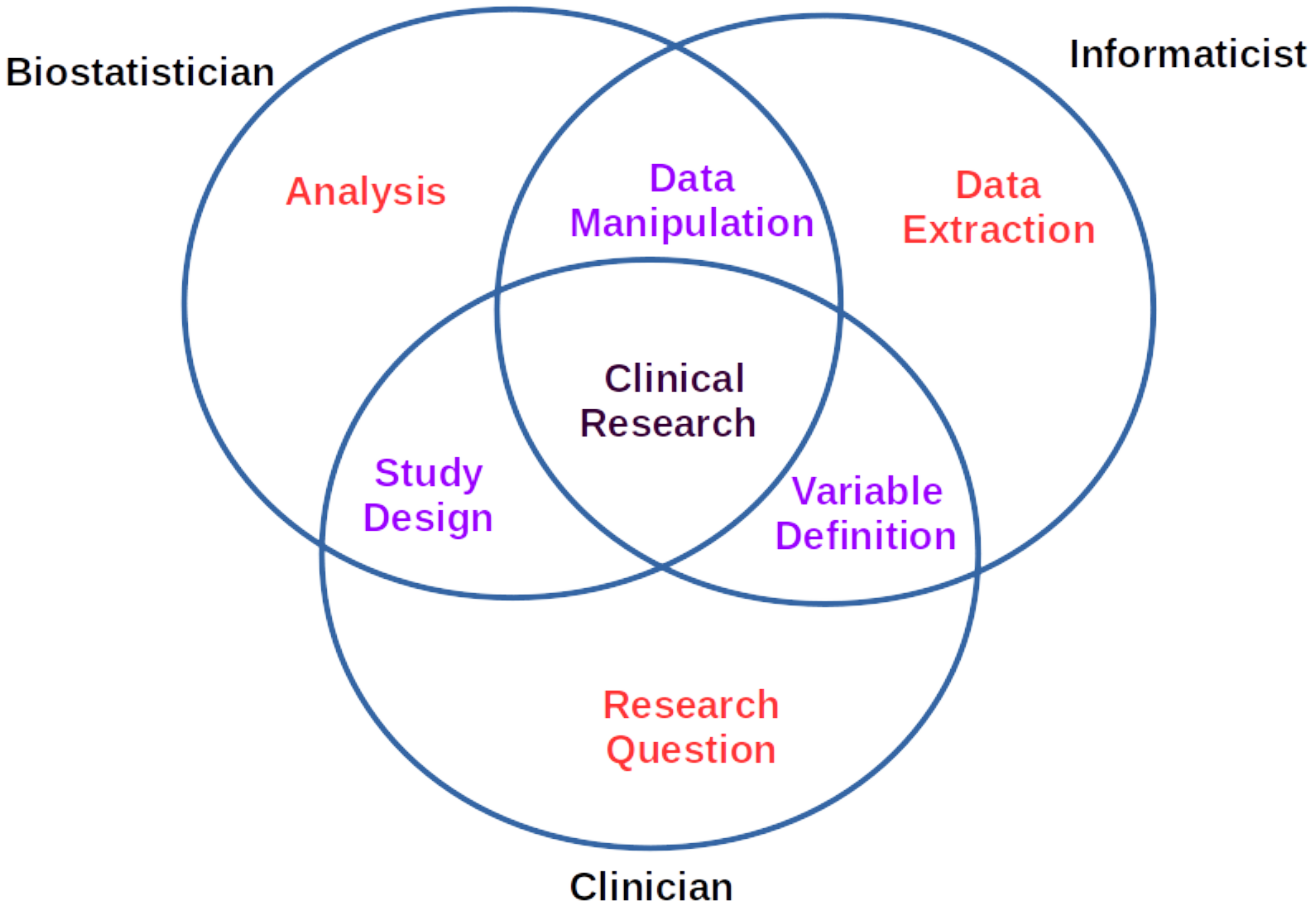
Accounting for Number of Encounters Attenuates Bias



Take Home

- Most analytic challenges arise based on how individuals seek care
- Need to be mindful of what may not be observed in EHR data
- Many challenges are controllable via the study & cohort design

Collaborative Clinical Research



References

- Goldstein BA, Phelan, M, Pagidipati,, NJ, & Peskoe, SB. How and When Informative Visit Processes Can Bias Inference when Using Electronic Health Records Data for Clinical Research. Submitted Journal of the American Medical Informatics Association.
- Bhavsar, NA., Gao, A., Phelan, M., Pagidipati, NJ., & Goldstein, BA. Value of Neighborhood Socioeconomic Status in Predicting Risk of Outcomes in Studies That Use Electronic Health Record Data. JAMA Network Open, 2018, 1(5): e182716.
- Phelan, M., Bhavsar, N.A., & Goldstein, B.A. Illustrating Informed Presence Bias in Electronic Health Records Data: How Patient Interactions with a Health System Can Impact Inference. EGEMS, 2017, 5(1): 22.
- Spratt, SE. Pereira, K., Granger, BB.,Batch, BC., Phelan, M., Pencina, M., Miranda, ML., Boulware, E., Lucas, JE., Nelson, CL., Neely, B., Goldstein, BA., Barth, P., Richesson, RL., Riley, IL., Corsino, L., McPeek Hinz, ER., Rusincovitch, S., Green, J., Barton, AB., and the DDC Phenotype Group., Assessing Electronic Health Record Phenotypes against Gold-Standard Diagnostic Criteria for Diabetes Mellitus. Journal of the Medical Informatics Association, 2017, 24, e121-e128.
- Goldstein B.A., Bhavasr, N.A., Phelan, M., & Pencina, M.J. Controlling for informed presence bias due to the number of health encounters in an Electronic Health Record. American Journal of Epidemiology, 2016, 184(11): 847-855.
- Richesson RL, Rusincovitch SA, Wixted D, Batch BC, Feinglos MN, Miranda ML, Hammond WE, Califf RM, & Spratt SE. A comparison of phenotype definitions for diabetes mellitus. Journal of the American Medical Informatics Association, 2013, (e2):e319-26.

Collaborators

- Nrupen Bhavsar
- Matt Phelan
- Sarah Peskoe
- Neha Pagidipati