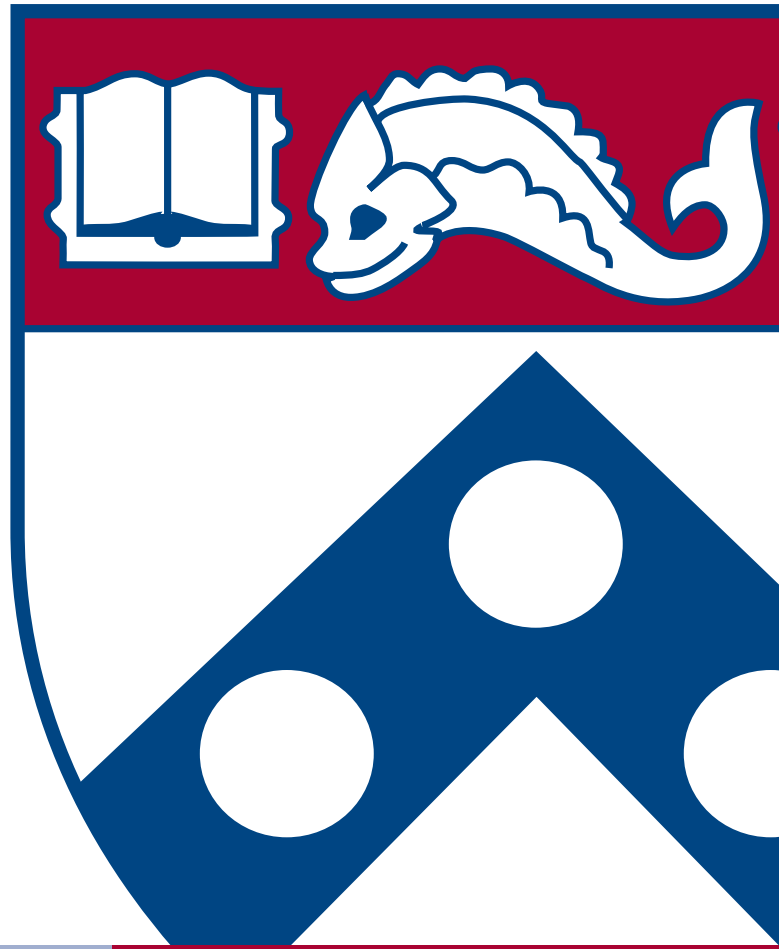
EHR-based Phenotyping in the Presence of Patient-Driven Observation Processes

Rebecca Hubbard

Dept. of Biostatistics, Epidemiology
& Informatics
University of Pennsylvania
rhubb@upenn.edu



October 23, 2018
3rd Seattle Symposium on
Healthcare Data Analytics

Acknowledgments

- Jing Huang
- Jinbo Chen
- Yong Chen
- Grace Choi
- Joanna Milton
- Arman Oganisian

This work was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1511-32666).
All statements in this presentation, including its findings and conclusions, are solely those of the authors and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.

Data provenance

- Data provenance refers to the process by which data come to be captured in the EHR
- Unlike data from a designed study, the data capture process in EHR-based studies is entirely outside the control (and often awareness) of the researcher
- Challenging aspects of data provenance for research include
  - Availability, type, and amount of data varies across patients
  - Clinical practices including frequency of visits, data that are recorded, tests that are ordered, etc may vary across clinics

Phenotype estimation using EHR data

- Phenotype = collection of characteristics describing a patient
- Motivated by lack of gold-standard for many patient characteristics of interest
- Need ways to deduce characteristics that are not explicitly recorded
- The complexities of data provenance create challenges for phenotyping
  - Patient-driven observation: Different data available for each patient, availability of data may be related to phenotype

Objectives

- Discuss challenges and alternative approaches for EHR-based phenotyping
- Propose a latent phenotyping model accounting for patient-driven observation
- Apply to the setting of T2DM using data from the PEDSnet federation, collection of children's hospitals participating in PCORI-funded network

Rule-based Phenotyping

- Most of the existing literature on EHR-derived phenotyping relies on "clinical decision rules"
- Algorithm based on clinical knowledge of the phenotype and coding practices
  - Simple or complex
  - Including one data element or many
  - May include a time component
- May incorporate structured data as well as unstructured data, often via NLP

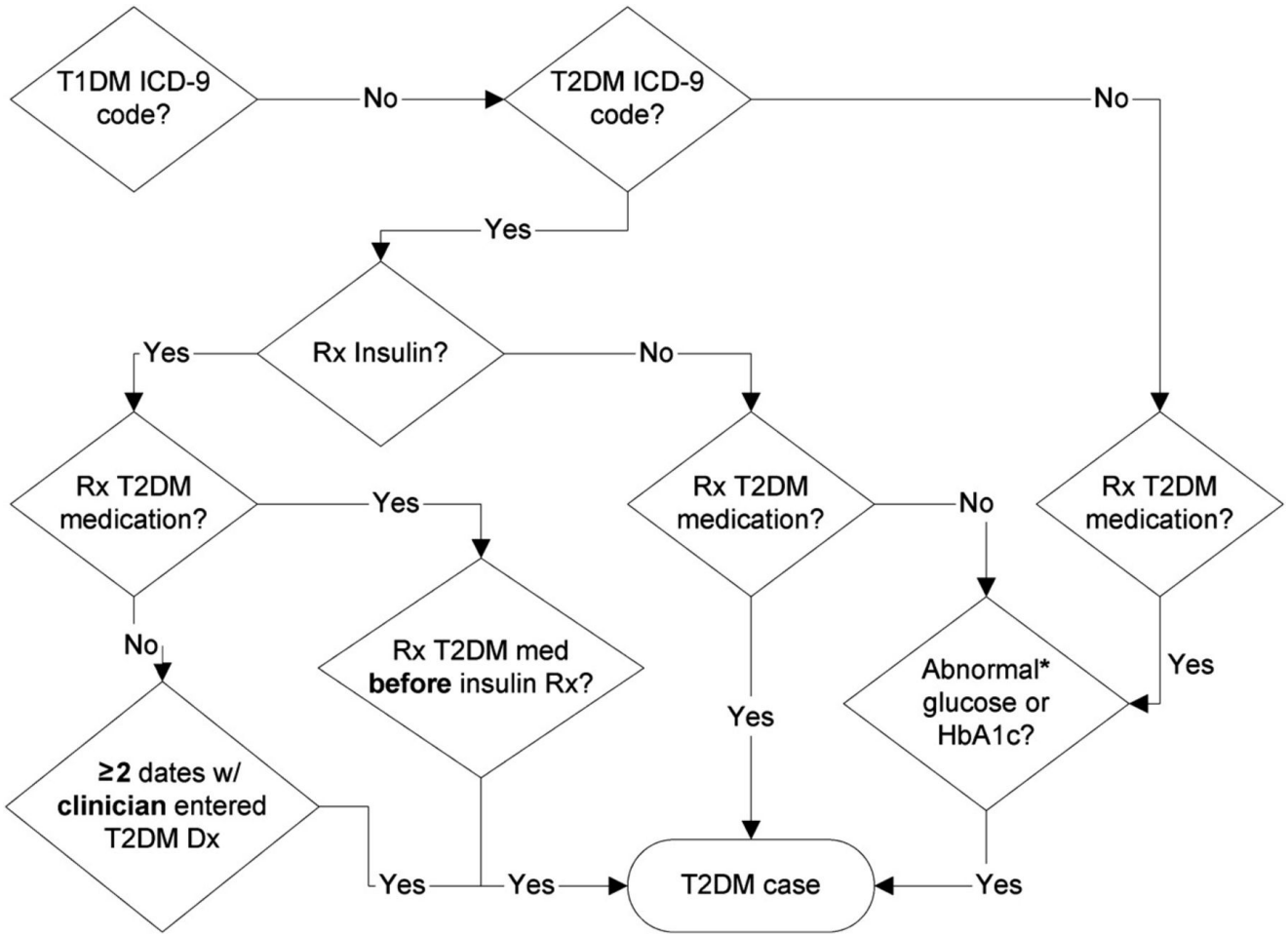
Typical process for EHR-based phenotype development

- Clinical experts develop a list of potential variables
  - May include condition of interest, symptoms, co-morbidities, common treatments
- Translate list into corresponding structured codes (e.g., ICD-9/10, SNOMED, CPT)
- Extract all occurrences of these codes from structured data
- Apply NLP to unstructured (narrative text) data
- Evaluate performance relative to gold-standard from manual chart review

Example: Rule-based Phenotyping for T2DM

Variable type	Examples	Format
Diabetes diagnosis	• T2DM • T1DM • DM NOS	ICD-9/10 codes
Medications	• Insulin • Metformin	Prescribing data
Co-morbidities	• PCOS • Obesity	ICD-9/10 codes
Biomarkers	• Glucose • HbA1c	Procedure codes for test administration; numerical results

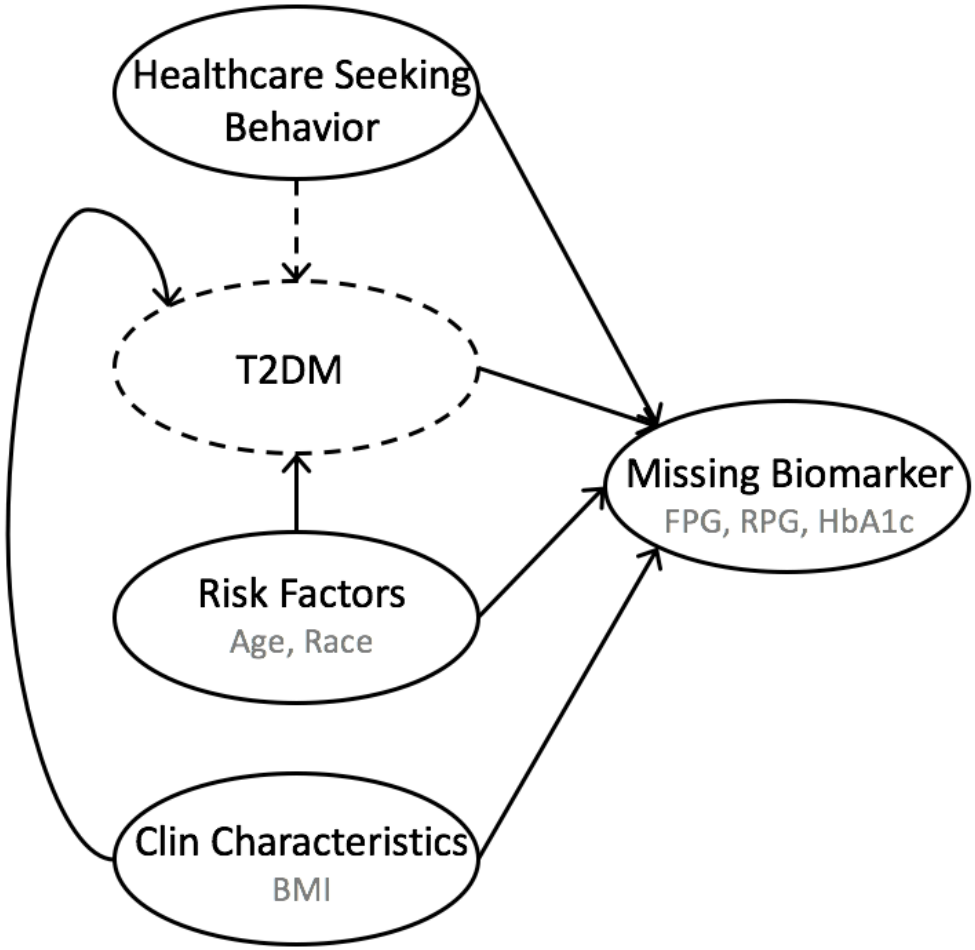
Example: T2DM Rule



Kho et al. *J Am Med Inform Assoc* 2012;19:212-218

MNAR missingness mechanism

- Missingness likely depends on underlying T2DM status directly
- Risk factors may influence missingness through T2DM (symptoms) or directly (screening)
- Patients' interaction with the healthcare system also affects observation process
- Example of patient-driven observation



A latent phenotype model

- As an alternative to rule-based phenotyping, we proposed a latent variable approach
- Assume each patient has an unobserved true phenotype Y_i
- Observable characteristics X_i (biomarkers, codes, medications) arise from distributions conditional on Y_i, $f(X_i|Y_i = k)$
- Missingness in biomarkers also incorporated as an observable characteristic conditional on Y_i, $f(R_i|Y_i = k)$
- Y_i arises from Bernoulli(θ_i)
- Estimates of $\theta_i|X_i$ can be used as continuous measures of predicted probability of phenotype

Hubbard et al. 2018. A Bayesian latent class approach for EHR-based phenotyping. *Statistics in Medicine*. doi:10.1002/sim.7953.

Application to PEDSnet data

- We applied this approach to an EHR-derived data set from two PEDSnet sites
- Children age 10-18 years, at least two clinical encounters between 2001-2017 separated by at least 3 years
- On at least one occasion BMI z-score in excess of the 95th percentile for age and sex
- Cohort consisted of 32,553 children from site A and 24,342 children from site B

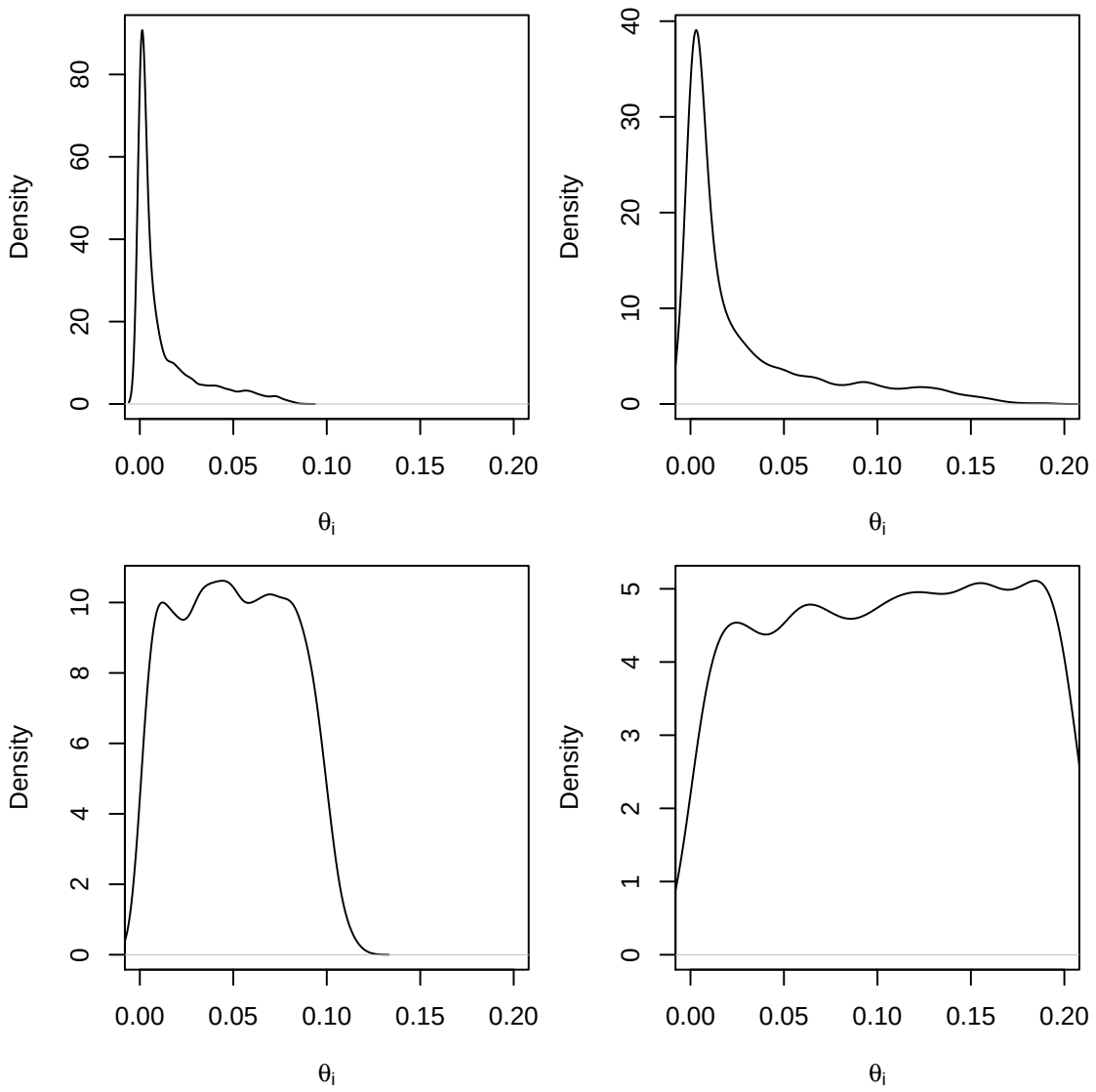
T2DM Predictors in PEDSnet cohort

	Site A	**Site B**
	N = 32,553	N = 24,342
	Mean (SD)	**Mean (SD)**
Random Glucose	95.0 (35.0)	101.8 (44.5)
Hemoglobin A1c	5.8 (1.2)	6.0 (1.4)
	N (%)	**N (%)**
Endocrinologist	2,411 (7.4)	4,617 (19.0)
Metformin	357 (1.1)	1,460 (6.0)
Insulin	360 (1.1)	691 (2.8)
T1D Codes	408 (1.3)	787 (3.2)
T2D Codes	164 (0.5)	365 (1.5)
Missing glucose	6,382 (19.6)	8,204 (33.7)
Missing HbA1c	29,057 (89.3)	18,630 (76.5)
eMERGE T2DM	111 (0.3)	207 (0.9)

Posterior means and CIs for model parameters

	Site A		Site B	
	Posterior Mean	95% CI	Posterior Mean	95% CI
Mean shift in glucose	135.24	(131.21, 139.25)	141.24	(138.87, 143.59)
T2DM code sensitivity	0.20	(0.16, 0.24)	0.26	(0.23, 0.29)
T2DM code specificity	1.00	(1.00, 1.00)	0.99	(0.99, 0.99)
Endocrinologist code sensitivity	0.95	(0.93, 0.97)	0.98	(0.97, 0.99)
Endocrinologist code specificity	**0.94**	**(0.94, 0.94)**	**0.84**	**(0.83, 0.84)**
OR missing glucose	**0.38**	**(0.31, 0.46)**	**0.20**	**(0.17, 0.23)**

Posterior density for T2DM



Conclusions and next steps

- Phenotyping is a fundamental first-step in EHR-based research
- Efforts should be made to improve phenotypes
  - Consider routine practice for how patients are treated and how frequently
  - Consider heterogeneity in clinical practice
  - Incorporate information on intensity of interaction with healthcare system
- A useful feature of our proposed approach is that it provides information on the predicted phenotype and a measure of its uncertainty
- Approaches are currently in development to improve incorporation of imperfect phenotypes into subsequent analyses

Thank you!
Questions?