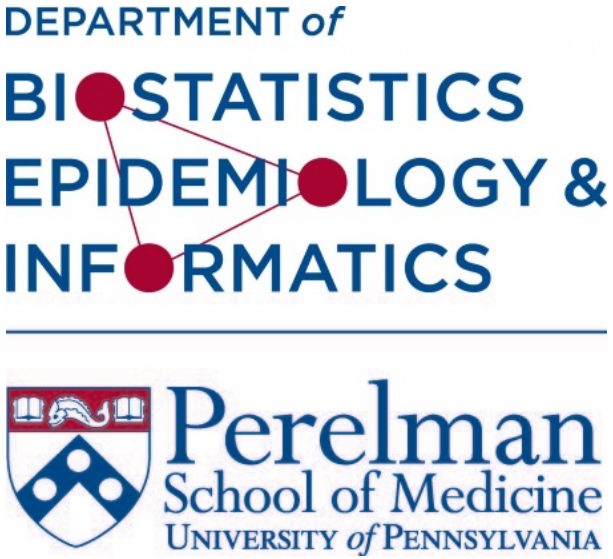
Fit for Use? The Role of Statistics in Transforming Electronic Health Records into Clinical Evidence

Rebecca Hubbard, PhD

rhubb@upenn.edu
https://www.med.upenn.edu/ehr-stats/

September 24, 2019

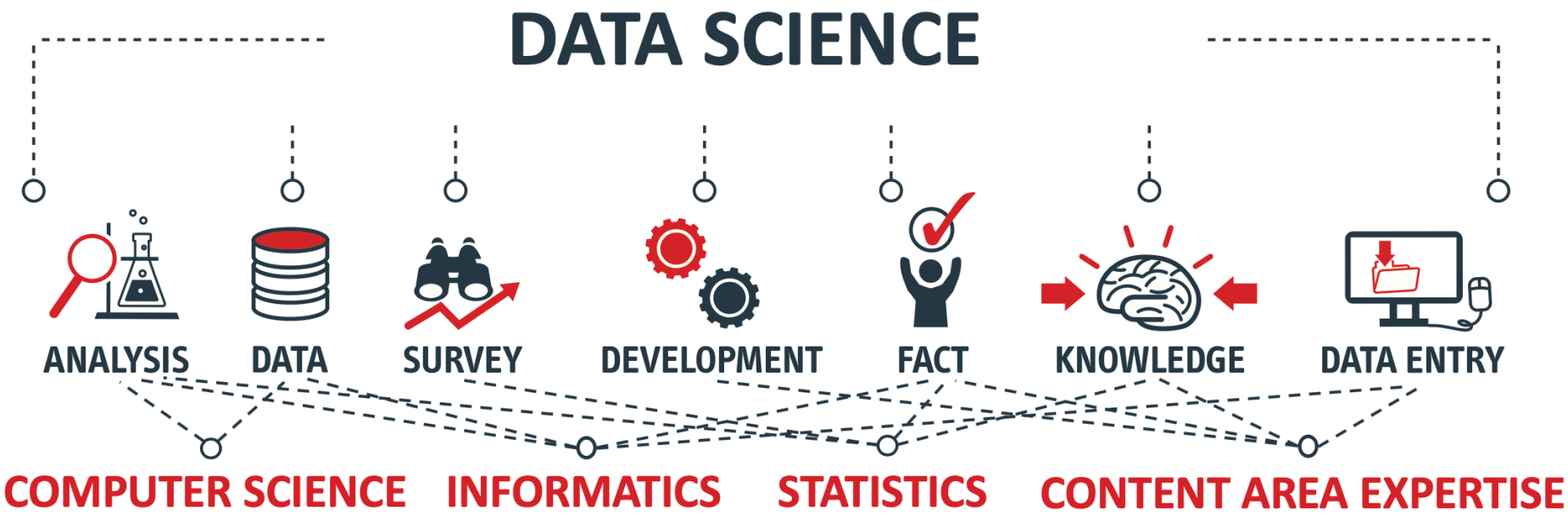
ASA Biopharmaceutical Section
Regulatory-Industry Statistics Workshop



Deriving Biomedical Knowledge from Real World Data

- Enormous medical research potential associated with study of RWD
- Clinical Informatics = "a body of knowledge, methods and theories that focus on the effective use of information and knowledge to improve the quality, safety and cose-effectiveness of patient care as well as the health of both individuals and populatons" – Detmer and Shortliffe. 2014. *JAMA*. 11: 2067-2068.
- In the context of EHR-based research, informaticians have capitalized on deep knowledge of databases and EHR data structures to lead in this field
- Statisticians have been hesitant to engage due to the (many) imperfections of these data

What are data scientists made of?



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

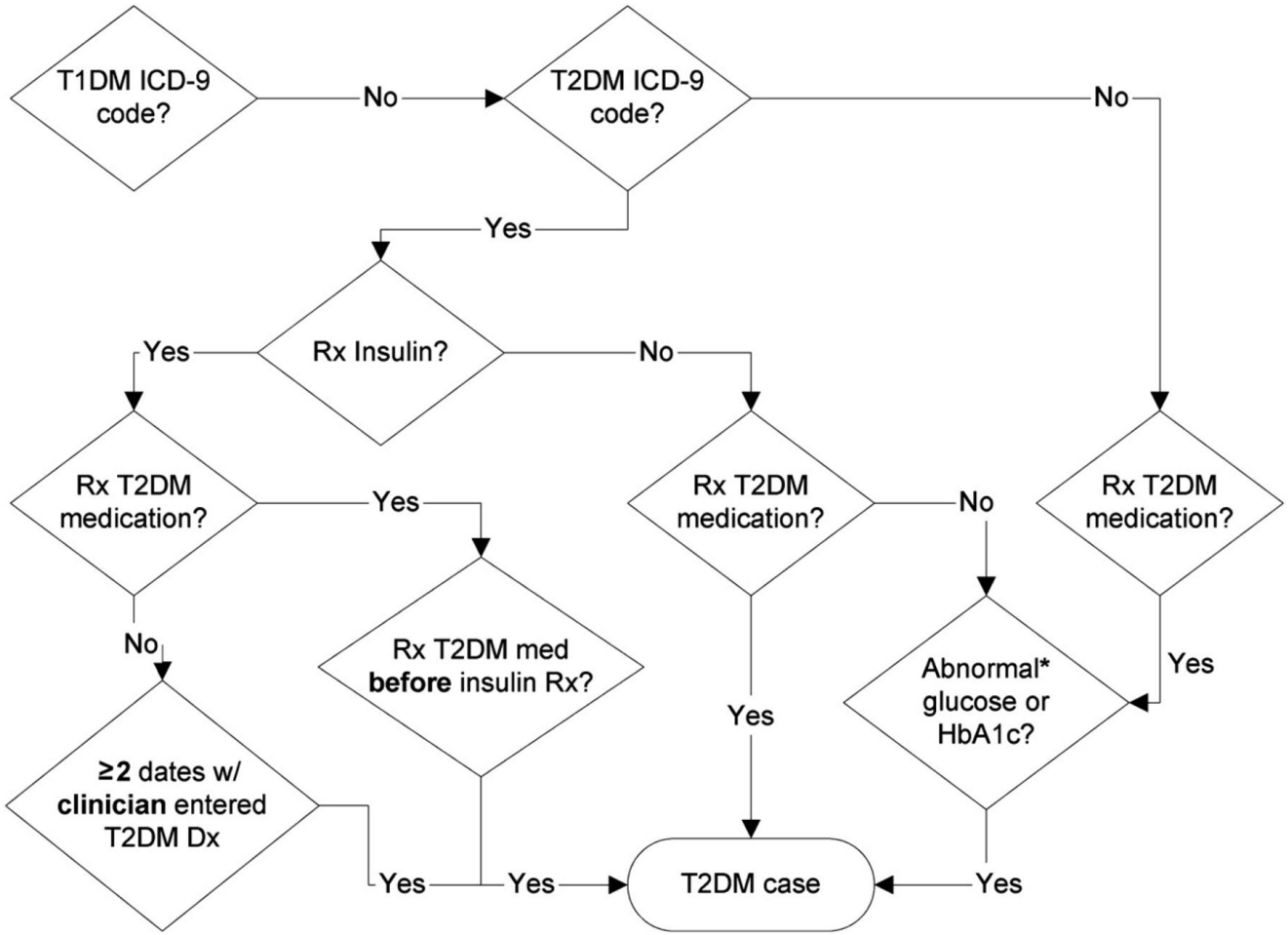
Phenotype estimation using EHR data

- 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

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 Natural Language Processing

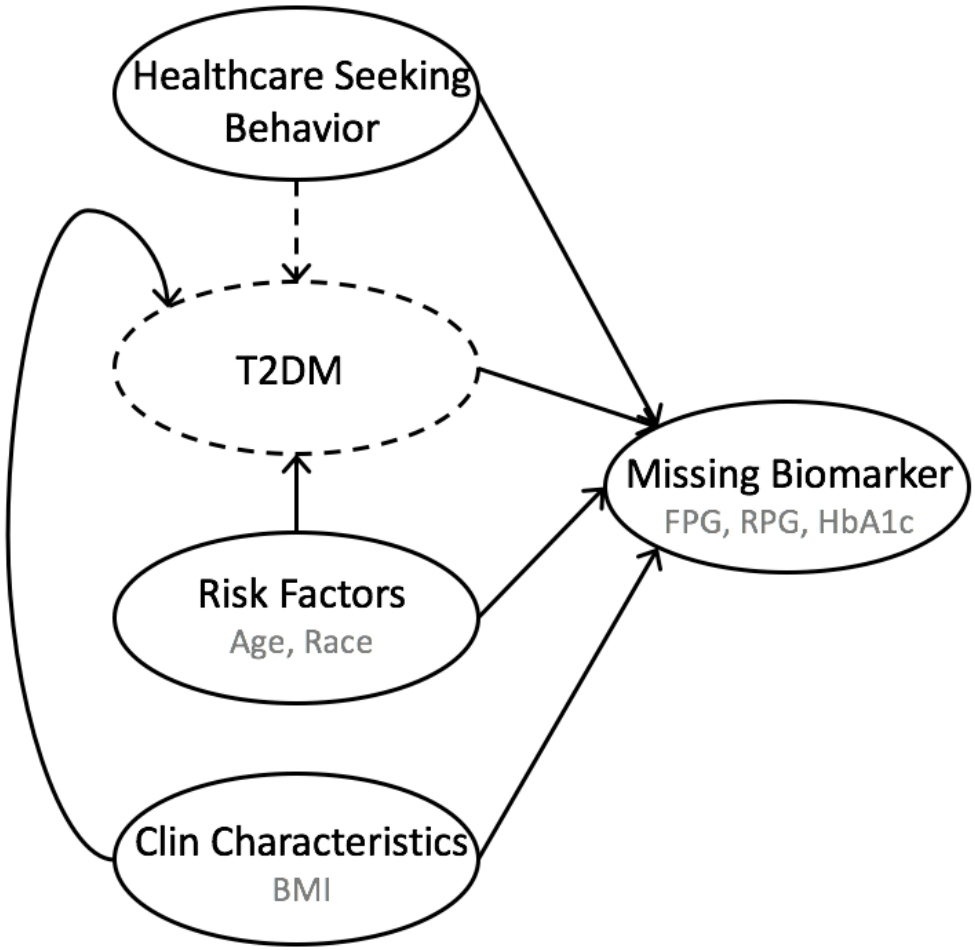
Example: eMERGE T2DM Rule



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

Unobserved true phenotype	$Y_i \sim \text{Bernoulli}(\theta_i)$
Observable features (e.g., codes, medications, biomarkers)	$\boldsymbol{X}_i \sim D(\boldsymbol{\mu}^X_{ik} \mid Y_i = k)$
Missingness in features	$\boldsymbol{R}_i \sim D(\boldsymbol{\mu}^{\boldsymbol{R}}_{ik} \mid Y_i = k)$
Priors for model parameters	$\pi(\theta_i)$, $\pi(\mu^X_{ik})$, etc

$$L(\theta_i) = \sum_{k=0,1} P(Y_i = k|\theta_i) \prod_{j=1}^{J} f(R_{ij}|Y_i = k) f(X_{ij}|Y_i = k)^{R_{ij}}$$

Posterior distribution for $\theta_i|\boldsymbol{X}_i, \boldsymbol{R}_i$ can be used as a measure of the phenotype

Hubbard et al. 2019. A Bayesian latent class approach for EHR-based phenotyping. *Statistics in Medicine*. doi:10.1002/sim.7953.

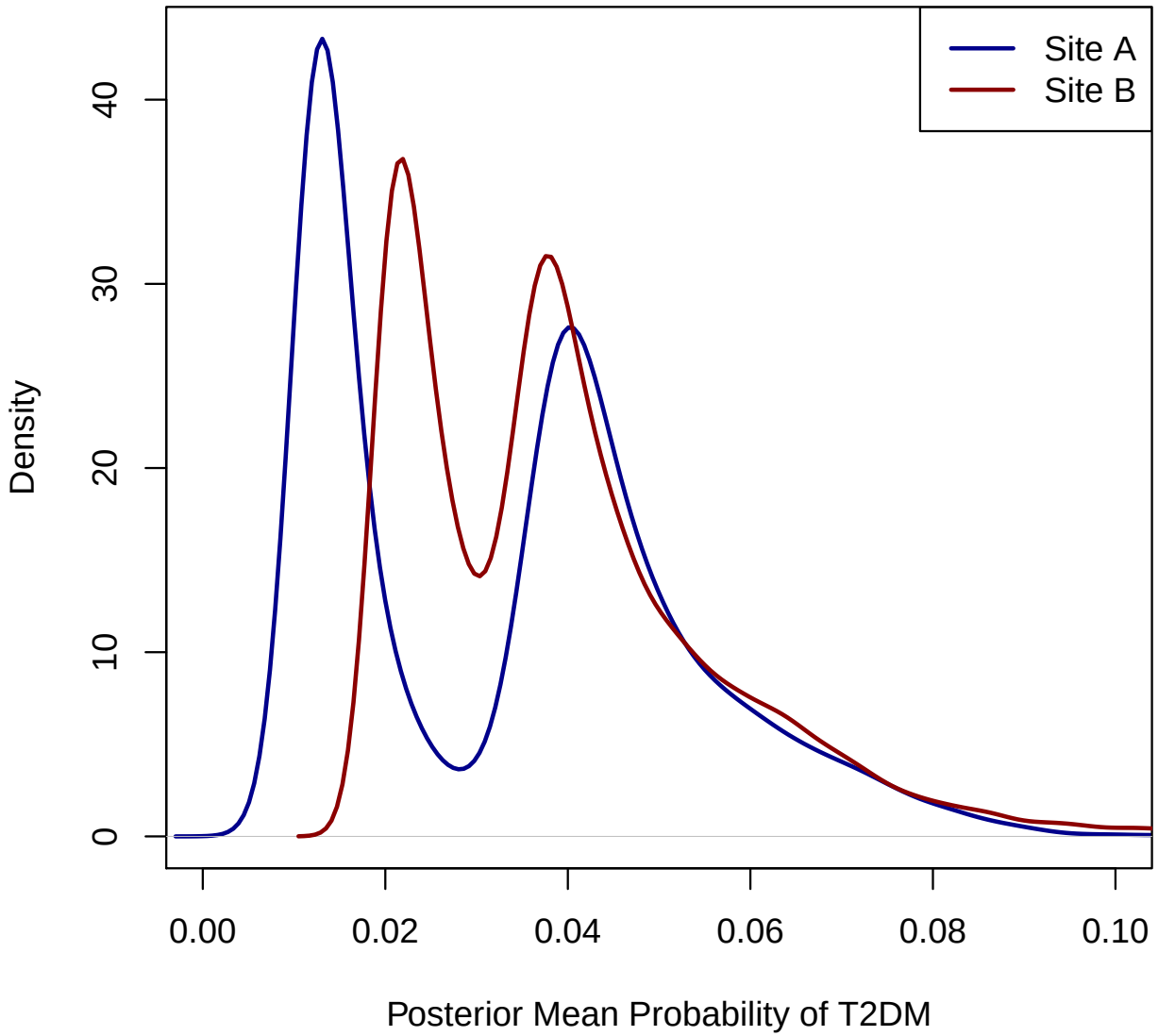
Application to PEDSnet data

- 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)

Distribution of posterior mean probabilities



Discussion

- Efforts should be made to improve phenotypes and incorporate knowledge of data provenance
  - Consider routine practice for how patients are treated and how frequently (requires collaboration with clinicians and coders)
  - Don't assume phenotypes are transportable
  - Use continuous probabilistic phenotypes, when available, rather than dichotomizing
- Data science takes a team: domain expertise, informatics, computer science, statistics
- Statistical science has the tools to address many of the challenges to valid inference presented by real world data

Acknowledgments

- Yong Chen
- Grace Choi
- Rui Duan
- Joanna Harton
- Jing Huang
- Arman Oganisian
- Jessie Tong

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.

Rebecca Hubbard

rhubb@upenn.edu
https://www.med.upenn.edu/ehr-stats/