Analysis of Big Healthcare Databases

Rebecca Hubbard, PhD

rhubb@upenn.edu
https://rhubb.github.io/

DEPARTMENT of
BIOSTATISTICS
EPIDEMIOLOGY &
INFORMATICS

Perelman
School of Medicine
University of Pennsylvania

Course Materials

All course materials can be downloaded from **https://rhubb.github.io/**

This includes:

- Slides
- Reference list
- Exercises
- Data sets used in exercises
- R code for exercise solutions

A little bit about me...

- Began my career at Kaiser Permanente Washington Health Research Institute (formerly Group Health RI), a public interest research group in an integrated health care system
- This gave me a lot of exposure to the opportunities of using EHR data for research but also the messiness and limitations of the data
- Since then I have worked with many administrative healthcare databases including
  - Medicare (public payer claims)
  - Optum Insight (private payer claims)
  - Kaiser Permanente (integrated health system and health plan)
  - U Penn (health system, tertiary care)
  - Flatiron Health (pooled data from community cancer care)
  - PEDSnet (multi-site network of children's hospitals)

EHR and the Real World

- EHR are one of the first “Real World Data” sources statisticians have gotten their hands dirty with
- The size of these data sets suggests enormous potential for learning about health and healthcare in real world settings.
  - Worldwide digital healthcare data is expected to reach 25 exabytes (10^{18} bytes) in 2020.
- But... with big data comes big responsibility
- That is, more data, more problems

Electronic Health Records

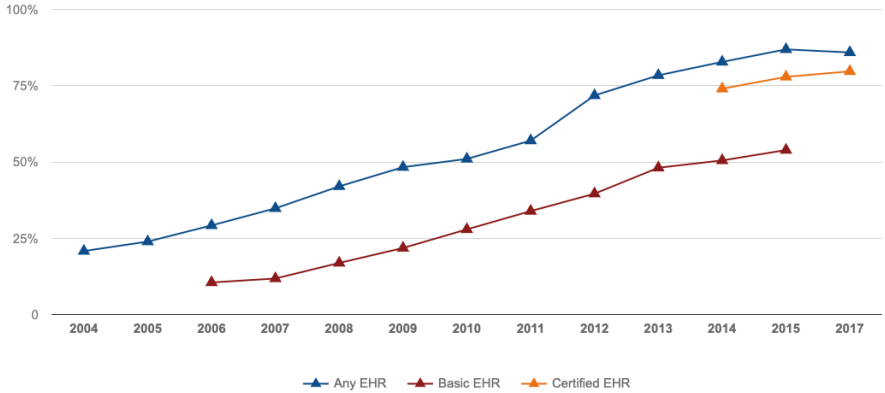
Definition

"An *Electronic Health Record (EHR)* is an electronic version of a patient's medical history, that is maintained by the provider over time, and may include all of the **key administrative clinical data** relevant to that persons care under a particular provider, including **demographics, progress notes, problems, medications, vital signs, past medical history, immunizations, laboratory data and radiology reports**."
– Centers for Medicare and Medicaid Services

Where do EHR data come from?

- Records from a single medical practice
- Records from a healthcare system including multiple practices
- Records from an integrated healthcare system including clinical data and claims
- Pooled data from multiple healthcare systems
- Regional or national databases in areas with a unified health system
- Claims data?
  - Medicare and Medicaid claims
  - Multi-payer claims databases

Physician EHR Adoption



Office of the National Coordinator for Health Information Technology. *Office-based Physician Electronic Health Record Adoption*, Health IT Quick-Stat #50. dashboard.healthit.gov/quickstats/pages/physician-ehr-adoption-trends.php. January 2019.

EHR Uptake

- Increased clinical use of EHRs has been driven largely by the Medicare and Medicaid EHR Incentive Program
- Under this program health care providers receiving reimbursement from Medicare are incentivized to adopt EHRs
- The objective of this program was to
  - Improve quality, safety and efficiency of health care and reduce health disparities
  - Engage patients and families in care
  - Improve care coordination
  - Improve population and public health
  - Ensure privacy and security of personal health information
- Regardless of whether these goals have been met or not, the practical implication for researchers is that large amounts of observational medical data are now available.

Objectives

- The objective of this short course is to present an overview of EHR data and methods for its analysis
- Focus is on **data structure and quality** particularly as they impact validity of research results
- Describe **strengths and limitations** of EHR data for research purposes
- Present **alternative statistical methods** to address challenges encountered in EHR data

Outline

What are the advantages of using EHR data for research?

- No need for patient recruitment or data collection
- Large sample size
- Diverse population
- Generalizability
- Potential for multi-site studies
- Cheap and “easy” access

What are the disadvantages of using EHR data for research?

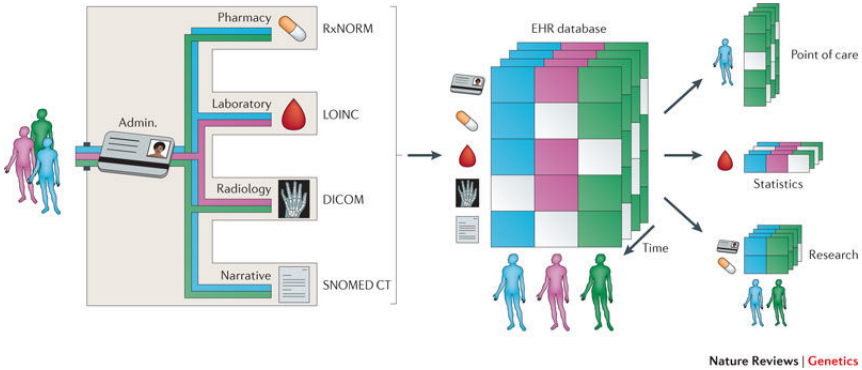
- Data quality may be poor (quantity vs quality tradeoff)
- Data collection is not systematic leading to complex missing data patterns
- Extracting data from text notes is challenging and error-prone
- Privacy protections (HIPAA) limit what data can be accessed and by whom

Dimensions of data quality

- **Completeness**: Is a truth about a patient present in the EHR?
- **Correctness**: Is an element that is present in the EHR true?
- **Concordance**: Is there agreement between elements in the EHR, or between the EHR and another data source?
- **Plausibility**: Does an element in the EHR makes sense in light of other knowledge about what that element is measuring?
- **Currency**: Is an element in the EHR a relevant representation of the patient state at a given point in time?

Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *Journal of the American Medical Informatics Association*. 2013;20(1):144-51.

Schematic of EHR data structure

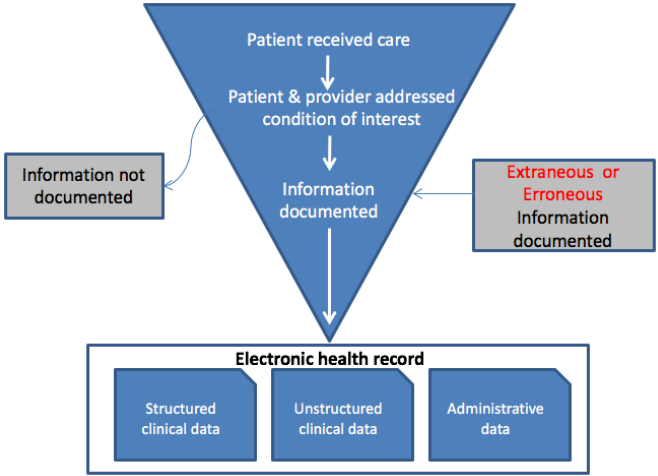


Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*. 2012;13(6):395-405.

From healthcare data to research data

- Conversion from raw healthcare data to raw research format is critical prior to analysis
- Typically need to work with a programmer or analyst with rights to access raw data, usually embedded in healthcare operations
- Refining research data prior to analysis limits volume of data, applies standardization to data elements, and reduces risks to privacy and confidentiality
- Adhere to “minimum necessary” rule
- When possible use de-identified data (no direct identifiers or HIPAA identifiers)

EHR data provenance



Structured data

- Many types of EHR data are available in *structured* form
- Structured data are standardized, pre-defined, computer readable data elements coded using a closed vocabulary
- For instance, procedure codes provide information on the specific health care procedures a patient has undergone
- Structured data are particularly useful for research because they can be readily manipulated and analyzed using statistical software
- They can also be combined across multiple healthcare systems using the same coding conventions.

Some common data formats

- International Classification of Disease-9/10 codes - diagnosis codes, issues with rule-outs, date of switch from ICD-9 to ICD-10
- CPT/HCPCS - procedure codes
- NDC - medication codes, may differ between healthcare systems, data may include codes for prescriptions (may not be filled or taken) or dispensings (may not be taken), does not capture OTC use

Unstructured data

- Unstructured data consists of health care providers' narrative clinical notes
- Takes the form of text which typically requires a human reader to understand and interpret
- The Health Story Project estimates that 1.2 billion clinical documents are produced in the US each year, of which 60% are in the form of unstructured notes
- Many data elements potentially of interest for research such as family history and patient behavioral risk factors may be embedded in text notes
- Extracting these data for research use is a challenge

Natural Language Processing

- Manually abstracting data from clinical notes is typically too labor intensive for large EHR-based studies which may include millions of patients
- Natural Language Processing (NLP) uses computerized algorithms to identify data elements of interest embedded in text notes
- NLP processes unstructured data (e.g., text notes) to identify “concepts” related to the factor of interest
- Standardized databases of health terminology such as the Unified Medical Language System link individual terms to unique biomedical concepts

Natural Language Processing in R

- R has many tools for conducting NLP
- This can be as simple as string manipulation
- Many functions for manipulating string are included in base R installation
  - tolower() - convert text to lower case
  - aspell() - correct spelling
  - substr() - extract substrings
  - regexpr() - regular expression matching
  - strsplit() - splits a string (e.g. sentence) into substrings (e.g. words)
- However, for more complex settings more sophisticated tools will be necessary

Natural Language Processing in R

- *Text Mining with R* available online at tidytextmining.com
- tm package includes more advanced tools for processing and manipulating text
  - Meyer D, Hornik K, Feinerer I. 2008. Text mining infrastructure in R. *Journal of Statistical Software*. 25(5):1-54.
  - Vignette available in package

Summary of EHR data

	ICD codes	CPT codes	Laboratory Data	Medication records	Clinical Documentation
Availability in EHR systems	Near-universal	Near-universal	Near-universal	Variable	Variable
Recall	Medium	Poor	Medium	Inpatient: High Outpatient: Variable	Medium
Precision	Medium	High	High	Inpatient: High Outpatient: Variable	Medium-High
Fragmentation effect	Medium	High	Medium-High	Medium	Low-Medium
Query method	Structured	Structured	Mostly structured	Structured, text queries, and NLP	NLP, text queries, and rarely structured
Strengths	-Easy to query -Serves as a good first pass of disease status	-Easy to query -High precision	-Value depends on test -High data validity	Can have high validity	Best record of what providers thought
Weaknesses	-Disease codes often used for screening when disease not a ctually present -Accuracy hindered by billing realities and clinic workflow	-Most susceptible to missing data errors (e.g., performed at another hospital) -Procedure receipt influenced by patient and payer factors external to disease process	-May need to aggregate different variations of the same data elements -Normal ranges and units may change over time	-Often need to interface inpatient and outpatient records -Medication records from outside providers not present -Medications prescribed not necessary taken	-Difficult to process automatically -Interpretation accuracy depends on assessment method -May suffer from significant "cut and paste" -Not universally available in EHRs -May be self-contradictory
Summary	Essential first element for electronic phenotyping	Helpful addition if relevant	Helpful addition if relevant	Useful for confirmation and a marker of severity	Useful for confirming common diagnoses or for finding rare ones

doi:10.1371/journal.pcbi.1002823.t001

Denny JC. Mining electronic health records in the genomics era. *PLoS Computational Biology.* 2012;8(12):e1002823.

PEDSnet: A multi-site network example

- PEDSnet: A PCORI-funded consortium of 8 children's hospitals
  - Includes data collected in routine clinical encounters for $\sim$5 million children
- Investigated pediatric Type 2 Diabetes Mellitus (T2DM) in a high risk cohort:
  - Children age 10-18 years, at least two clinical encounters between 2001-2017
  - On at least one occasion BMI z-score in excess of the 95th percentile for age and sex
  - Cohort consisted of 68,265 children

PEDSnet T2DM cohort

	Total	T2DM Codes or Biomarkers	
		Yes	No
	N = 68,265	N = 804	N = 67,461
	N (%)	**N (%)**	**N (%)**
Male	36836 (53.96)	221 (27.49)	36615 (54.28)
White	35740 (52.35)	371 (46.14)	35369 (52.43)
Endocrinologist	5338 (7.82)	510 (63.43)	4828 (7.16)
Metformin	764 (1.12)	675 (83.96)	89 (0.13)
Insulin	727 (1.06)	154 (19.15)	573 (0.85)
T2DM Codes	275 (0.4)	221 (27.49)	54 (0.08)
Any glucose measurement	11325 (16.59)	355 (44.15)	10970 (16.26)
Any HbA1c measurement	6031 (8.83)	397 (49.38)	5634 (8.35)
	Mean (SD)	**Mean (SD)**	**Mean (SD)**
Age	11.897 (2.498)	13.791 (2.581)	11.874 (2.488)
BMI z-score	2.015 (0.303)	2.269 (0.361)	2.012 (0.301)
Glucose	94.309 (32.511)	141.393 (104.471)	92.785 (27.438)
Hemoglobin A1c	5.786 (1.254)	6.934 (1.940)	5.705 (1.149)

Fit for Use

- “Fit for use” means that the data should be of appropriate quality for the use we intend to put them to.
- Almost all data sources are imperfect in one respect or another.
- The relevant question is whether they are good enough for the purposes we intend to use them for.
- For example, if we are interested in studying the association between practicing yoga and blood pressure, EHR data are probably not fit for this use.
  - No structured data elements capture practice of yoga
  - Text notes are unlikely to systematically records yoga

Fit for Use

- On the other hand, if we want to study rates of uptake of HPV testing for cervical cancer screening EHR data may be fit for this use.
  - Procedure codes capture HPV testing
  - Since a procedure code must be recorded in order for the provider to be reimbursed capture is likely to be relatively complete
- As we discuss EHR data and its analysis it is important to keep in mind that the relevance of a given issue will depend on the specific research question and the data needed to answer that question.

Exercises

- Throughout this course we will use a synthetic (i.e., fake) EHR-derived data set based on the structure of the PEDSnet dataset to explore some of the challenges of EHR data and strategies for dealing with them we will be discussing
- The PEDSnet synthetic data sets and R code for working with them are available at `https://rhubb.github.io`
- These data sets were generated by sampling from the distributions of data elements contained in the real PEDSnet data
- The result is a set of data sets that reflect the features of the PEDSnet data but avoid issues of privacy protection that accompany real EHR-derived data

Exercise 1

Outline

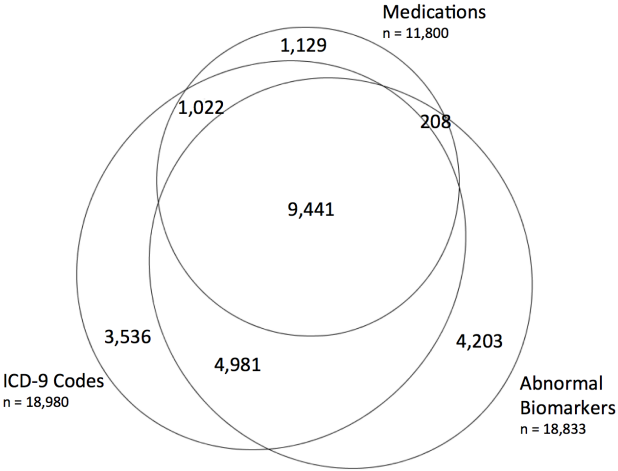
Rule-based Phenotyping

- Most of the existing literature on EHR-derived phenotyping relies on “clinical decision rules”
  - Simple or complex
  - Including one data element or many
  - May include a time component
- Algorithm based on clinical knowledge of the disease and coding practices
- May incorporate structured data as well as unstructured data, often via NLP
- Sometimes validated against gold-standard diagnosis data

Example: Rule-based Phenotyping for T2DM

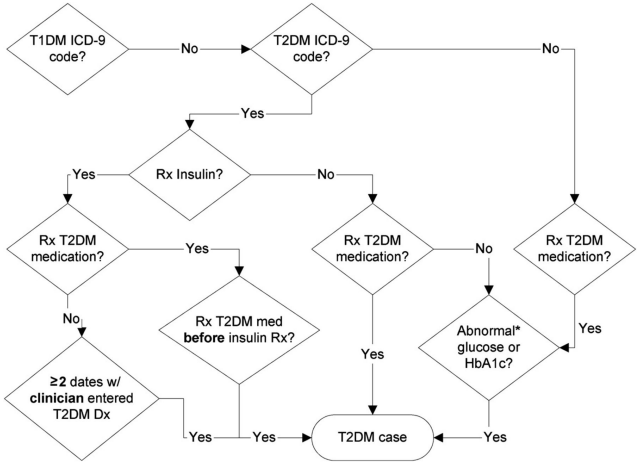
Variable type	Examples	Format
Diabetes diagnosis	• T2DM • T1DM • DM NOS	ICD-9/10 codes
Medications	• Insulin • Metformin	Prescribing data
Co-morbidities	• PCOS • Obesity	ICD-9/10 codes
Biomarkers	• Glucose • HbA1c	Procedure codes for test administration; numerical results

Agreement among TD2M variables



Adapted from Richesson et al. *J Am Med Inform Assoc* 2013;20:e319-e326.

Example: T2DM Rule



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

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, CPT)
- NLP experts map terms to UMLS concepts
- Extract all occurrences of demographics, codes of interest, biometric data, and laboratory test results from structured data
- Apply NLP to unstructured (narrative text) data

PEDSnet Example

age	race	zBMI	HbA1c	Glucose	T2DM	FamHx	Endo
11	0	1.658355	NA	NA	0	0	0
12	0	1.996588	NA	119.54555	0	0	0
14	1	2.057993	5.949531	98.69711	0	0	0
18	0	2.508225	5.137229	82.54253	0	0	0
11	1	1.820784	NA	NA	1	0	0
17	1	2.547955	5.622635	85.22707	0	0	0

Classification

- Once data have been extracted from the EHR a classification algorithm can be applied to the individual data elements to create a construct of interest
- Gold standard information for supervised learning approaches extracted via manual chart abstraction
- Classification approaches applied to EHR data range from the very simple to the very complex
  - Dichotomous classification based on presence/absence of data elements based on clinical judgment
  - Prediction modeling, e.g. CART, LASSO
  - Machine learning algorithms, e.g. random forests, neural networks
- Performance is typically evaluated based on PPV and NPV relative to gold standard
- Implications of low prevalence for PPV/NPV

Using validated phenotypes

- Ideally, only validated phenotypes should be used
- Validation requires manual chart abstraction and hence can be costly and slow
- Many validated phenotypes are available, for instance, via PheKb (https://phekb.org)
- However, be cautious about assuming that operating characteristics will be the same in your data set as they were in the derivation data set (i.e., lack of portability)

Exercise 2

Outline

Missing data in EHR

- Because EHR data are not collected according to a research protocol they will often be missing variables of interest
- While missing data are virtually ubiquitous in EHR-based studies, a critical first step to dealing with missingness is consideration of what constitutes a “complete” record
- Unlike a designed observational study, there is no prior specification of which data elements should be collected for a patient or when they should be collected
- Before we can quantify the magnitude of the missing data problem for a given study we need to define the data we wish to have
- Often useful to consider the data that we would have collected had we designed a study protocol and collected the data ourselves

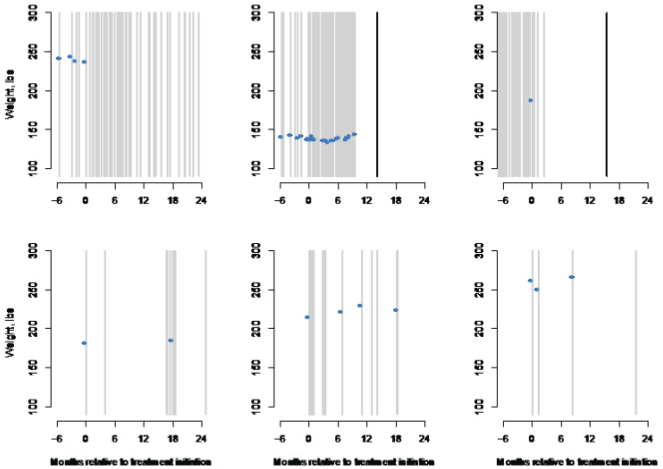
PEDSnet Example

- Objective: Study risk factors for T2DM in children
- Longitudinal study of time to T2DM diagnosis
- Covariates:
  - Time-varying measures of BMI, physical activity, diet, co-morbidities
  - Age at diagnosis of T2DM
- If this is our desired study objective and design, which of these data elements can be derived from the EHR and how much missingness will they have?

PEDSnet Example

- Patients assessed irregularly; must decide on a frequency of observation for BMI that is "good enough"
- Behavioral risk factors rarely collected, not in structured data; may be able to extract from notes with NLP but will be frequently missing
- Age at diagnosis determined based on application of a phenotyping algorithm; depending on algorithm may be missing for patients with missing biomarker data or infrequent clinical assessment

BMI data collection

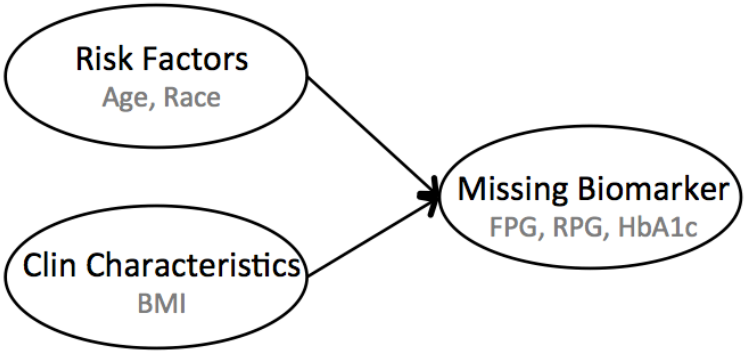


Haneuse and Daniels. 2016. *eGEMs* 4(1):16.

Missingness mechanisms

- Once we have decided on which data we need for our study we can evaluate missingness
- Next step is to consider causes of missing data
- Haneuse and Daniels recommend thinking about why data are *observed* rather than why data are missing
- Missingness mechanism will likely reflect interplay of patient risk factors and disease conditions, patient behavior, provider clinical practices, and healthcare system administrative practices (**data provenance**)

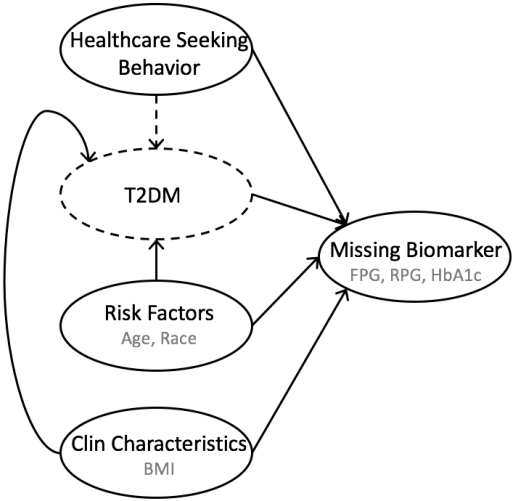
MAR missingness mechanism



- Patients with risk factors for T2DM more likely to be screened
- Risk factors more strongly associated with missingness for more definitive biomarkers (FPG, HbA1c)

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

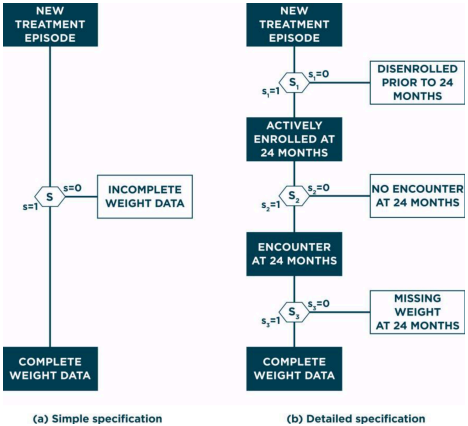


PEDSnet example: Missingness in glucose

	OR	95% CI	p
Age at baseline (years)	0.89	(0.88, 0.90)	<0.001
Baseline year	1.02	(1.02, 1.03)	<0.001
Male	1.14	(1.09, 1.19)	<0.001
Race			
Black	0.83	(0.79, 0.87)	<0.001
Native American	0.60	(0.33, 1.11)	0.106
Asian	0.90	(0.75, 1.08)	0.280
Other	0.57	(0.52, 0.62)	<0.001
Unknown	1.39	(1.25, 1.54)	<0.001
Ethnicity			
Hispanic	0.81	(0.75, 0.88)	<0.001
Endocrinologist visit	0.23	(0.21, 0.24)	<0.001
Metformin	0.30	(0.26, 0.35)	<0.001
Insulin	1.10	(0.86, 1.40)	0.454
BMI	0.69	(0.66, 0.73)	<0.001
T1D codes	0.91	(0.73, 1.14)	0.404
T2D codes	0.59	(0.45, 0.77)	<0.001

Missingness modules

- Considering missingness mechanism as a series of conditional steps may help in assessment of MAR assumption (Haneuse & Daniels, 2016)



Analysis strategies in the presence of missingness

- Under MAR mechanisms can use multiple imputation (MI) or inverse probability weighting (IPW)
- Many software packages available for implementation
- In Multivariate Imputation via Chained Equations (MICE) a separate regression model is specified for each variable with missing observations
- Missing data in each variable are sequentially filled in and subsequently used in regression models for other variables
- This process is iterated until parameter estimates are stable
- Predictions are then generated from the final set of models for all missing observations

MICE

- MICE is convenient for use with EHR data because regression models for each variable can allow for different variable types and can include different predictors
- However, the process of model specification can be quite laborious
- Additionally, MICE is somewhat ad hoc in that the set of conditional models for each variable may not correspond to a joint model for all variables
- Computationally intensive for large EHR samples
- Available in many software packages including the `mice` package for R

Loss to follow-up in EHR

- A challenging aspect of longitudinal studies using EHR is that we may not know when patients have left the healthcare system
- Claims databases provide an indicator of enrollment that can be used to censor patients who disenroll
- A variety of ad hoc approaches have been proposed including
  - Censoring patients at a fixed timepoint (e.g. 1 year) after last clinical encounter
  - Restrict cohort to patients with some level of interaction with healthcare system (Lin KJ, et al. Identifying patients with high data completeness to improve validity of comparative effectiveness research in electronic health records data. *Clinical Pharmacology & Therapeutics*. 2018;103(5):899-905.)
- Care must be taken to avoid immortal time bias

A case study of missing outcome data

Carrigan G, et al. 2019. An evaluation of the impact of missing deaths on overall survival analyses of advanced non–small cell lung cancer patients conducted in an electronic health records database. *Pharmacoepidemiology and drug safety*. doi:10.1002/pds.4758

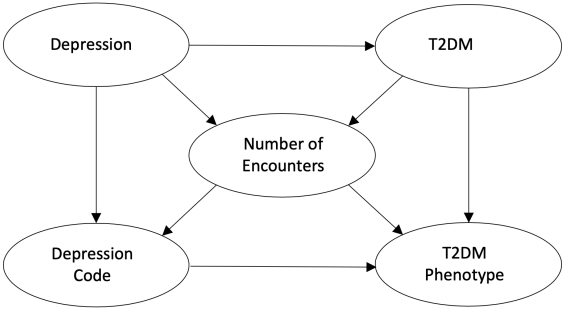
- Used data from the Flatiron Health database of community oncology centers and gold-standard death data from National Death Index to investigate effect of missing information on mortality on hazard ratio estimates
- Missing outcomes substantially inflated estimates of median survival time but had little effect on hazard ratios
- However, using EHR data as external control arm resulted in significant bias
- Implications for between-site comparisons if loss to follow-up patterns differ

Biased sampling in EHR data

- Complex observation patterns also arise in terms of the number of observations per subject in EHR data
- In EHR data, some members of the target population are more frequently observed than others
  - Co-morbidity and health outcomes: patients with co-morbidities that require regular monitoring (e.g. diabetes, kidney disease) have more frequent contact with healthcare system; capture health outcomes data more frequently and accurately
  - Screening test performance: patients experiencing symptoms of the disease of interest more likely to come in for screening tests
- If intensity of interaction with the healthcare system is related to the disease of interest, this results in an informative observation scheme, violating the assumptions of many standard statistical methods

Informative observation processes

- Intensity of healthcare utilization can be considered a marker of health
- In this case, patients with many visits may be systematically different from those with few
- One way to deal with this is to condition on number of encounters
- This has been shown to effectively account for informative observation processes in some cases, but can induce M-bias



Goldstein BA et al. Controlling for informed presence bias due to the number of health encounters in an electronic health record. *Am J Epidemiol*. 2016 Dec 1;184(11):847-55.

Exercises 3 and 4

Outline

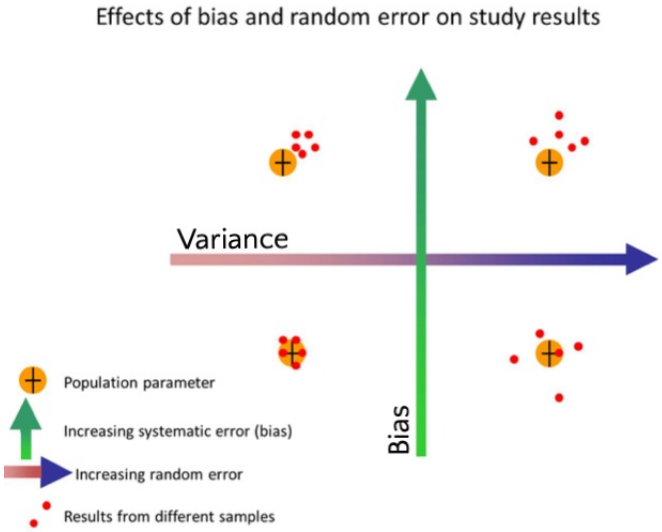
Issues arising in the analysis of EHR-derived phenotypes

- Regardless of the approach to phenotyping, some residual error will typically remain
- Statistical methods for misclassified outcomes can be adapted to this context
- Some additional challenges in the context of EHR-based research arise due to limited access to validation data
- Accuracy parameters for phenotypes may also exhibit lack of transportability
- We will discuss some alternative approaches to these challenges

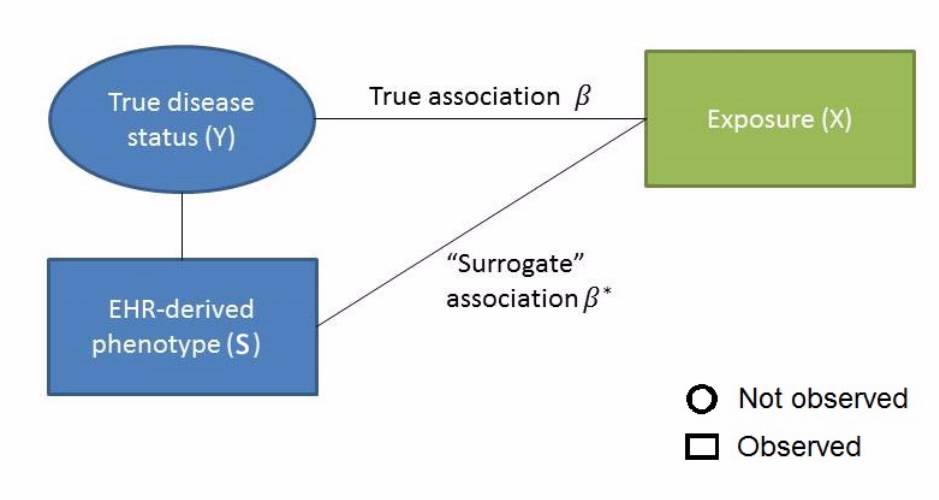
What can we do about phenotyping error?

- Phenotyping error can lead to substantial bias and inflated type I error
- How can we obtain unbiased estimates in association analyses using EHR-derived phenotypes?
- Challenges in the EHR setting
  - Validation data are costly to obtain and in many cases completely unavailable
  - Phenotyping accuracy is often unknown and may vary widely between derivation data set and other data sets

The bias variance tradeoff in big data



Data Structure



Classic approach to outcome misclassification

- One binary predictor (X)
- Misclassified binary outcome (Y)
- Known sensitivity (θ) and specificity (ϕ); $0 < \theta, \phi < 1$

	Classified as	
	Diseased	Not diseased
Exposed	a	b
Not exposed	c	d

Classic approach to outcome misclassification

- Naive: $\widehat{OR}_{standard} = (ad)/(bc)$
- Misclassification adjusted: $\widehat{OR} = \frac{a/(a+b)-(1-\phi)}{c/(c+d)-(1-\phi)} \times \frac{\theta-c/(c+d)}{\theta-a/(a+b)}$
- Note that

$$\widehat{OR} > \widehat{OR}_{standard} \text{ if } \widehat{OR}_{standard} > 1$$
$$\widehat{OR} < \widehat{OR}_{standard} \text{ if } \widehat{OR}_{standard} < 1$$

Magder LS, Hughes JP. 1997. Logistic regression when the outcome is measured with uncertainty. *Am J Epidemiol.*146(2):195-203.

Extension to logistic regression

Definition

Non-differential misclassification
Let Y = true outcome, S = surrogate outcome
$S \perp X|Y$**, or equivalently**
$\theta = P(S = 1|Y = 1, X) = P(S = 1|Y = 1)$**,**
$\phi = P(S = 0|Y = 0, X) = P(S = 0|Y = 0)$

Extension to logistic regression

- Assume non-differential misclassification
- Let $P(Y_i = 1) = expit(\beta^T X_i)$ then by Bayes rules
  - $\hat{P}(Y_i = 1|S_i = 1) = \frac{\theta expit(\beta^T X_i)}{\theta expit(\beta^T X_i)+(1-\phi)(1-expit(\beta^T X_i))}$
  - $\hat{P}(Y_i = 1|S_i = 0) = \frac{(1-\theta) expit(\beta^T X_i)}{(1-\theta) expit(\beta^T X_i)+\phi(1-expit(\beta^T X_i))}$
- An EM algorithm to estimate β
  1. Perform weighted logistic regression, each subject included as both diseased and non-diseased with weights $\hat{P}(Y_i = 1|S = k)$
  2. Update weights using new values for $\hat{\beta}$
  3. Return to (1)

An approach for predicted probabilities

- Increasingly, phenotyping uses statistical or machine learning approaches that provide probabilistic phenotypes, $\hat{p}$
- Sinnott et al. 2014 developed a bias correction approach for analyses using these predicted probabilities as outcomes
- Suppose we wish to estimate the association between a phenotype, Y, and exposure, Z adjusting for confounders W

$$g(P(Y = 1|Z, W)) = \alpha + \beta Z + \gamma W.$$

- Let $f(\hat{p}) = (\hat{p} - \mu_0)/(\mu_1 - \mu_0)$, where $\mu_k = E(\hat{p}|Y = k)$
- Sinnott et al. showed that regressing $f(\hat{p})$ on Z and W provides unbiased estimates for regression coefficients.

Sinnott et al. 2014. Improving the power of genetic association tests with imperfect phenotype derived from electronic medical records. *Human Genetics*. 133:1369-82.

A simple bias correction for risk differences

- In the context of risk difference regression in which the link function, $g(\cdot)$, is the identity link, this approach gives rise to a very simple bias correction

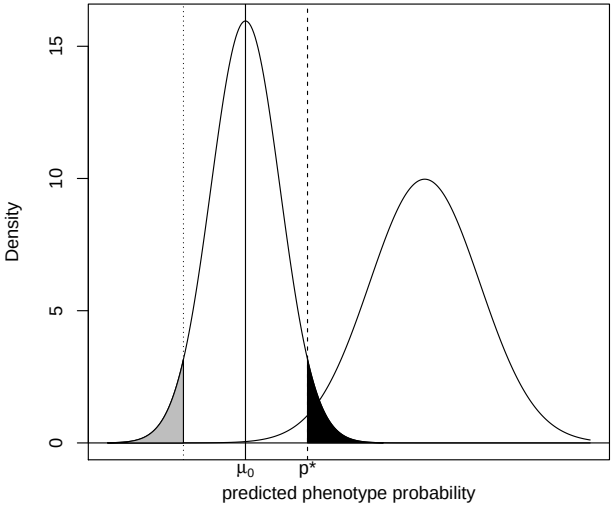
$$\begin{aligned} E(f(\hat{p})|Z,W) &= \alpha + \beta Z + \gamma W \\ E[(\hat{p}-\mu_0)/(\mu_1-\mu_0)|Z,W] &= \alpha + \beta Z + \gamma W \\ E[\hat{p}|Z,W] &= \alpha^* + (\mu_1-\mu_0)(\beta Z + \gamma W) \\ E[\hat{p}|Z,W] &= \alpha^* + \beta^* Z + \gamma^* W \end{aligned}$$

- Therefore, $\hat{\beta} = \frac{\hat{\beta}^*}{\mu_1-\mu_0}$ is unbiased for β
- By using a Taylor series expansion to linearize $g(\cdot)$ bias correction formulas can be obtained for other link functions including log and logistic.

One additional complication

- Unfortunately, in the EHR context μ_0 and μ_1 will only be available in data sets with validation data
- In the data set initially used to develop the phenotype this will be straightforward to calculate by taking the mean of $\hat{p}$ among cases and controls
- In data sets without validation data we typically have access to published validation results, typically including a proposed cutpoint, p^*, along with sensitivity and specificity for the dichotomized phenotype
- Using this information we can obtain estimates $\hat{\mu}_0$ and $\hat{\mu}_1$

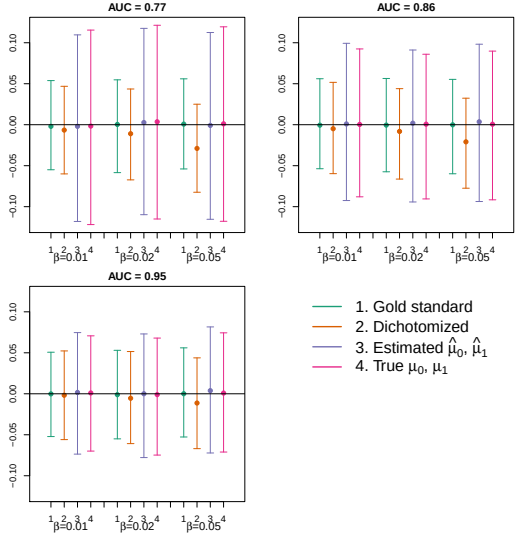
Estimating μ_0 without validation data



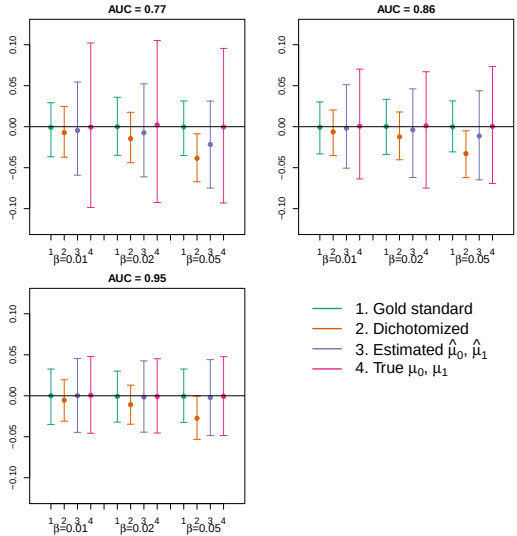
Simulation study design

- Compared
  1. Gold standard true phenotype
  2. Dichotomized phenotype based on predicted probability
  3. Bias correction using estimated $\hat{\mu}_0$ and $\hat{\mu}_1$
  4. Bias correction using true μ_0 and μ_1
- Varying: AUC of $\hat{p}$, strength of effect (β), prevalence of Y

Bias: Prevalence = 0.5



Bias: Prevalence = 0.1



Exercises 5 and 6

Outline

Exploring EHR data

- If you are interested in exploring EHR data there are a number of sources available online
- **MIMIC III**: publically available data on 40,000 critical care patients
  - https://mimic.physionet.org
  - Requires DUA for full data access
- **healthdata.gov**: open access data from the US government
  - Includes data from many sources including Medicare claims
  - Some data sets are limited to aggregate data
  - Medicare PUF include individual-level data but not suitable for research

Concluding thoughts

- Due to financial incentives and operational efficiencies, EHR will become the dominant mode of clinical/administrative documentation of health encounters
- This creates a vast research resource but also requires knowledge of its complexities to use appropriately
- A key component of data science is expert knowledge about data sources
- To effectively use EHR data we (statisticians) must be willing to learn about where these data come from and how they are used clinically/administratively
- We wouldn't analyze observational data without reading the protocol!

Recommendations

- Engaging with clinicians, coders, informaticians allows us to
  - Understand data quality
  - Make smart choices about when and how EHR data can be used
  - Identify appropriate methods to mitigate limitations
  - Develop new statistical methods to fill gaps in available methodology
- EHR data can be messy but don't despair!
- Staying engaged in the research process from data extraction through analysis, interpretation, and reporting of results ensures higher quality research and gives us a seat at the table to help improve processes for the future

Acknowledgments

- Jinbo Chen
- Yong Chen
- Grace Choi
- Jessica Chubak
- Joanna Harton
- Jing Huang
- Arman Oganisian
- Jessie Tong
- Jianqiao Wang
- Weiwei Zhu

ASA Council of Chapters

Rebecca Hubbard, PhD

rhubb@upenn.edu

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