

# Leveraging electronic health records for data science

Jesse Gronsbell

Department of Statistical Sciences  
University of Toronto



CMS Winter Meeting 2022

*Stochastic Systems, Probability, and Other Mathematical Aspects of  
Data Science*

## Roadmap for today

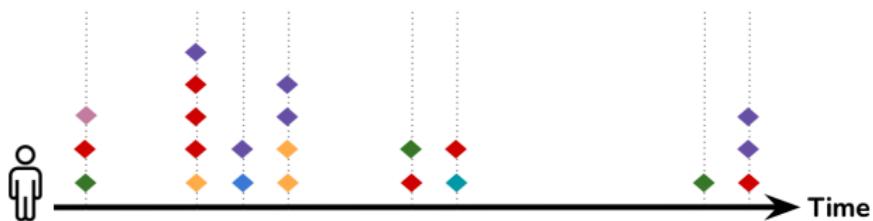
---

- Background on electronic health records (EHRs)
- Opportunities & challenges for data science
- Precise model evaluation with scarce labeled data
- Future directions

# What is an Electronic Health Record (EHR)?

An electronic record of a patient's interactions with a healthcare system

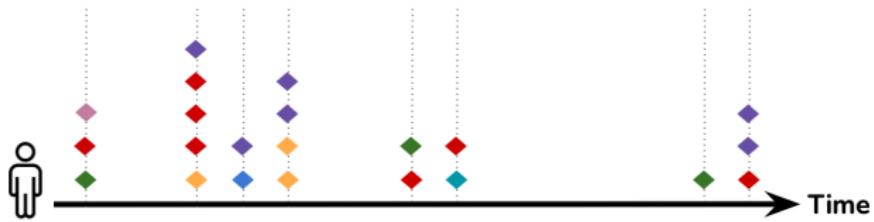
- ◆ Demographics
- ◆ Diagnoses
- ◆ Lab tests
- ◆ Medications
- ◆ Procedures
- ◆ Vital signs
- ◆ Clinical notes



# What is an Electronic Health Record (EHR)?

An electronic record of a patient's interactions with a healthcare system

- ◆ Demographics
- ◆ Diagnoses
- ◆ Lab tests
- ◆ Medications
- ◆ Procedures
- ◆ Vital signs
- ◆ Clinical notes



EHR data is a byproduct of clinical care

## The blessing: EHR data is extensive

---

**Big** Longitudinal records on large populations

**Detailed** Information on numerous fields

**Representative** Real-world patients

↑ **Available** Increasing EHR adoption worldwide

# The opportunity: Learn from EHR data



THE WALL STREET JOURNAL

THE FUTURE OF EVERYTHING | DATA

## Medical Records Data Offers Doctors Hope of Better Patient Care

*Healthcare professionals are beginning to tap the treasure trove of information locked in electronic health records to treat people in real time*

*Healthcare professionals are beginning to tap the treasure trove of information locked in electronic health records to treat people in real time*

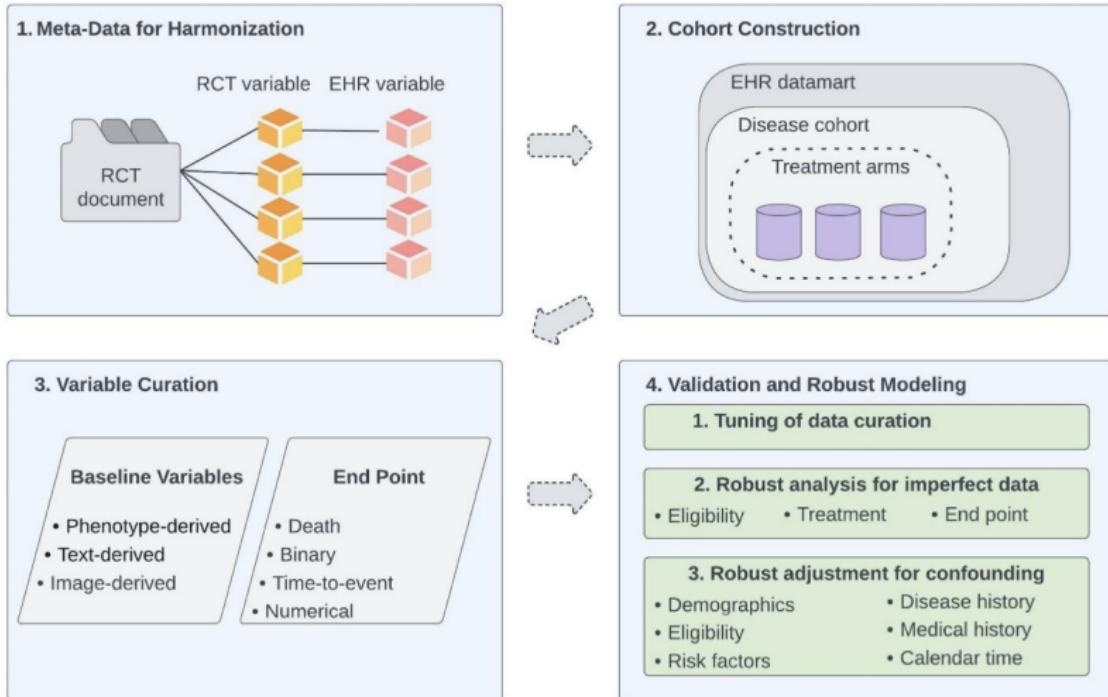
# Data science at the bedside: A “Green Button”

“**a green patients like mine button** as a tool in the EHR would both **support patient care decisions** in the absence of published evidence and, as a byproduct, quantify and **prioritize unanswered clinical questions** for EHR-enabled randomization at the point of care ”

*Longhurst et al 2014*



# Data science at the bench: Real-world evidence



## The challenge: EHR data is not research ready

---

**EHRs do not have explicit information on phenotypes**

# The challenge: EHR data is not research ready

---

**EHRs do not have explicit information on phenotypes**

Phenotype: patient characteristics inferred from EHRs

- Presence of a disease
- Disease severity or subtype
- Time of disease onset
- Disease progression
- Treatment response
- ...

# Phenotypes are the foundation of EHR research

- Presence of a disease
- Disease severity or subtype
- Time of disease onset
- Disease progression
- Treatment response
- ...

} Identify and characterize  
the population of interest  
Targets of risk prediction,  
causal inference, ...

→ Cohort identification

eg. Who are the patients with rheumatoid arthritis?

# Phenotypes are the foundation of EHR research

- **Presence of a disease**
- Disease severity or subtype
- Time of disease onset
- Disease progression
- Treatment response
- ...

} Identify and characterize  
the population of interest

**Targets of risk prediction,  
causal inference, ...**

→ Risk prediction

eg. How do low density lipoprotein risk alleles affect the risk of cardiovascular disease among rheumatoid arthritis patients?

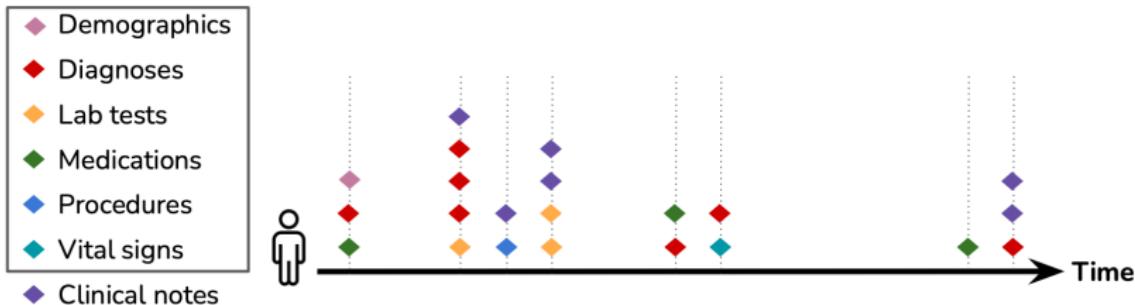
# Why is phenotyping challenging?

---

"Health data is like crude oil. It is useless unless it is refined."

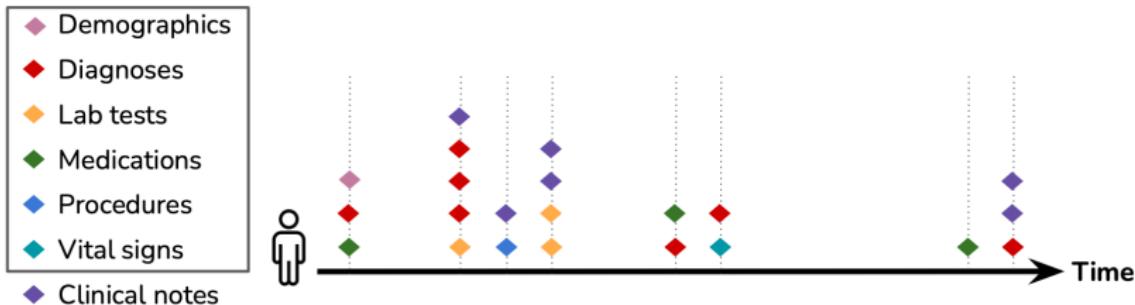
*Leo Anthony Celi*

# The two flavors of EHR data



**1. Structured data:** Easy to extract, but lacks context

# The two flavors of EHR data

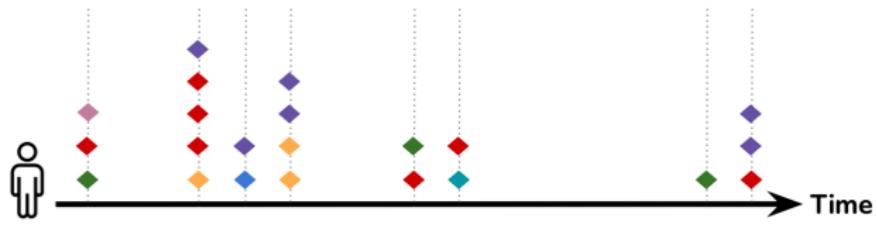


**1. Structured data:** Easy to extract, but lacks context

eg. diagnosis code  $\not\rightarrow$  disease diagnosis  
complex diagnosis, upcoding, temporal shift, etc.

# The two flavors of EHR data

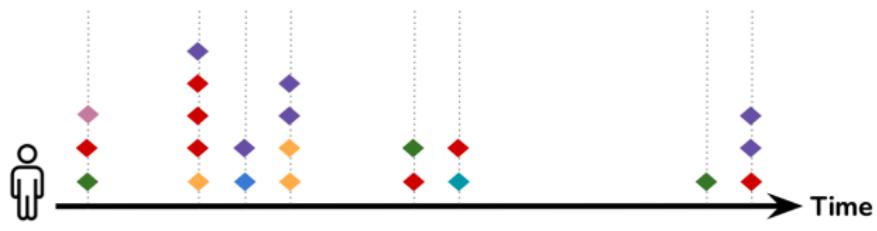
- ◆ Demographics
- ◆ Diagnoses
- ◆ Lab tests
- ◆ Medications
- ◆ Procedures
- ◆ Vital signs
- ◆ Clinical notes



**2. Unstructured data:** Rich information, but requires NLP

# The two flavors of EHR data

- Demographics
- Diagnoses
- Lab tests
- Medications
- Procedures
- Vital signs
- Clinical notes



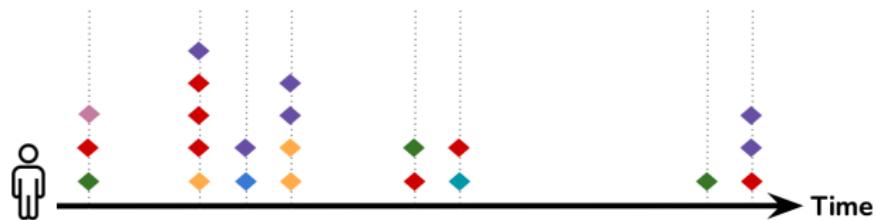
**2. Unstructured data:** Rich information, but requires NLP

clinical terms → concept unique identifier (CUI)

eg. “Rheumatoid Arthritis”, “RA” = C000387

# The two flavors of EHR data

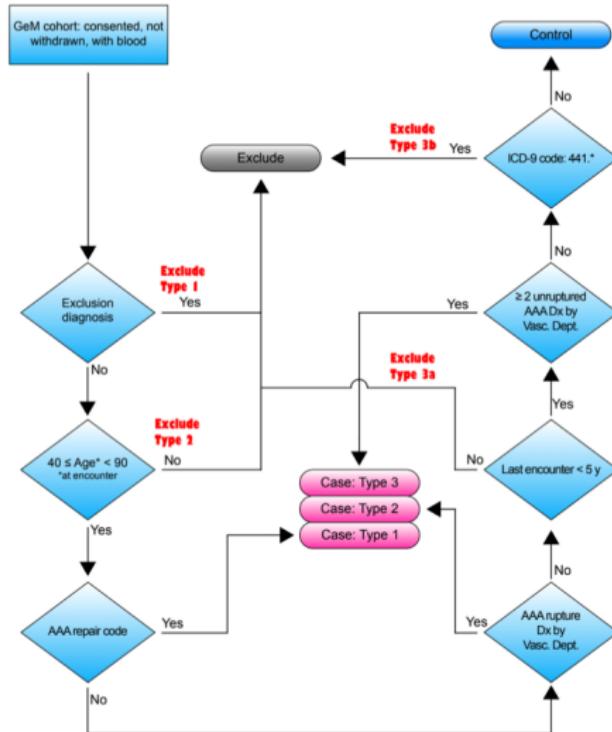
- ◆ Demographics
- ◆ Diagnoses
- ◆ Lab tests
- ◆ Medications
- ◆ Procedures
- ◆ Vital signs
- ◆ Clinical notes



phenotype  $\approx$  structured + unstructured data

# Most phenotyping algorithms are rule-based

## Abdominal Aortic Aneurysm



# Moving forward: ML-based phenotyping algorithms

*Journal of the American Medical Informatics Association*, 00(0), 2022, 1–15

<https://doi.org/10.1093/jamia/ocac216>



Review

---

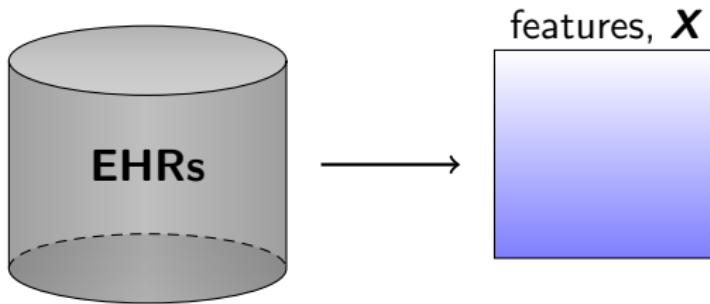
Review

## **Machine learning approaches for electronic health records phenotyping: a methodical review**

Siyue Yang<sup>1</sup>, Paul Varghese<sup>2</sup>, Ellen Stephenson <sup>3</sup>, Karen Tu<sup>3</sup>, and  
Jessica Gronsbell<sup>1,3,4</sup>

# Building a ML-based phenotyping algorithm

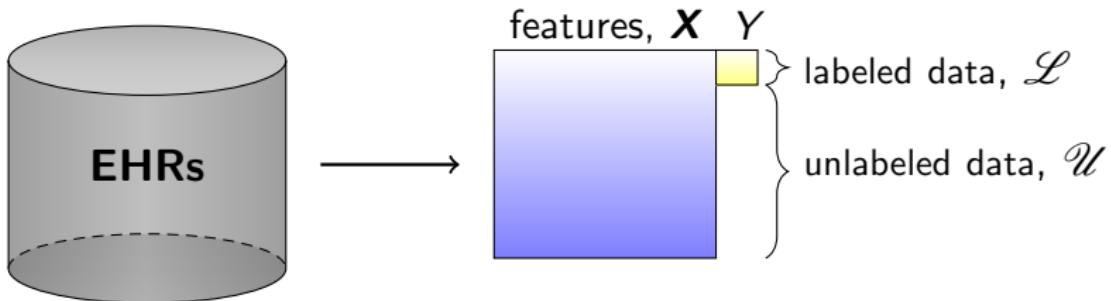
**Step 1:** Extract relevant features,  $X$ , for patients of interest



- Structured data: diagnosis codes, medications, labs
- Unstructured data: positive mentions of procedures, labs, medications, and signs and symptoms

# Building a ML-based phenotyping algorithm

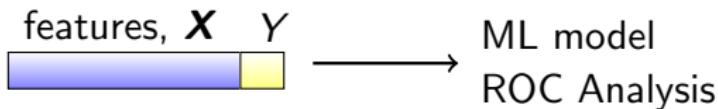
**Step 2:** Manually review records to label phenotype status,  $Y$



# Building a ML-based phenotyping algorithm

**Step 3:** Use  $\mathcal{L}$  to estimate and evaluate a model

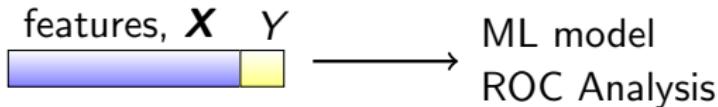
$$Y \sim g(\mathbf{X}; \boldsymbol{\theta}) \rightarrow \hat{Y}(\mathbf{X}) \sim g(\mathbf{X}; \hat{\boldsymbol{\theta}})$$



# Building a ML-based phenotyping algorithm

**Step 3:** Use  $\mathcal{L}$  to estimate and evaluate a model

$$Y \sim g(\mathbf{X}; \boldsymbol{\theta}) \rightarrow \hat{Y}(\mathbf{X}) \sim g(\mathbf{X}; \hat{\boldsymbol{\theta}})$$

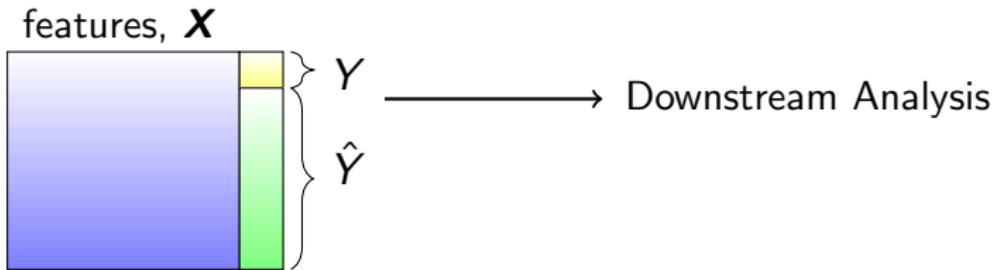


1<sup>st</sup> algorithm for rheumatoid arthritis

- ALASSO logistic regression ( $n = 500$ ): AUC = 0.95

# Building a ML-based phenotyping algorithm

**Step 4:** Project the model predictions to  $\mathcal{U}$



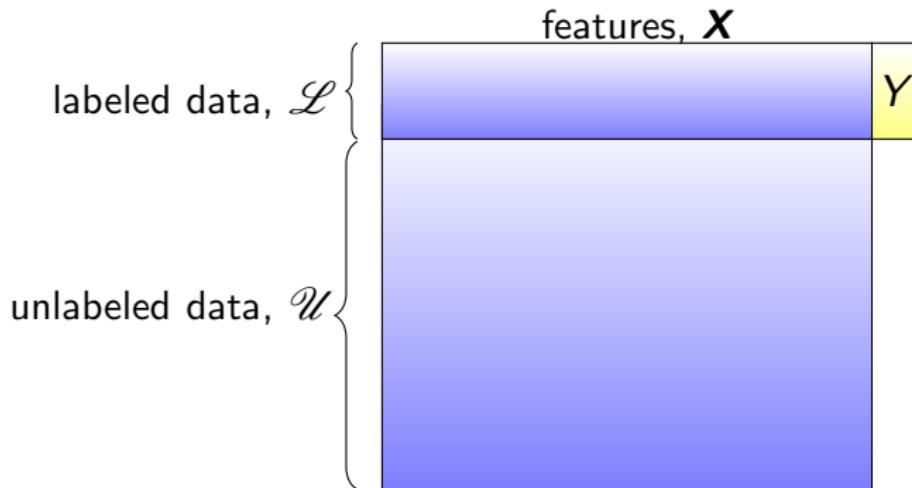
# Building a ML-based phenotyping algorithm

1. Extract relevant features,  $X$ , for patients of interest
2. Manually review records to label phenotype status,  $Y$
3. Use  $\mathcal{L}$  to estimate and evaluate a model
4. Project the model predictions to  $\mathcal{U}$

“It took 20 MD/PhDs 2 years to do this.”

# How can we make phenotyping more efficient?

→ Leverage all of the information that is available to us



Make use of both the labeled **and** unlabeled data

## How can we make phenotyping more efficient?

---

Make use of the labeled **and** unlabeled data for:

1. Model estimation (eg. regression model)
2. Model evaluation (eg. ROC parameters)

# How can we make phenotyping more efficient?

Make use of the labeled **and** unlabeled data for:

1. Model estimation (eg. regression model)
2. Model evaluation (eg. ROC parameters)

Gronsbell, J., Minnier, J., Yu, S., Liao, K. and Cai, T., 2019. [Automated feature selection of predictors in electronic medical records data.](#)  
Biometrics, 75(1), pp.268-277.

Yu, S., Ma, Y., Gronsbell, J., Cai, T., Ananthakrishnan, A.N., Gainer, V.S., Churchill, S.E., Szolovits, P., Murphy, S.N., Kohane, I.S. and Liao, K.P., 2018. [Enabling phenotypic big data with PheNorm.](#) Journal of the American Medical Informatics Association, 25(1), pp.54-60.

## How can we make phenotyping more efficient?

---

Make use of the labeled **and** unlabeled data for:

1. Model estimation (eg. regression model)
2. Model evaluation (eg. ROC parameters)

# How can we make phenotyping more efficient?

Make use of the labeled **and** unlabeled data for:

1. Model estimation (eg. regression model)
2. Model evaluation (eg. ROC parameters)

Gronsbell, J. and Cai, T., 2018. [Semi-supervised approaches to efficient evaluation of model prediction performance](#). Journal of the Royal Statistical Society: Series B, Statistical Methodology, 80(3), pp.579-594.

Gronsbell, J., Liu, M., Tian, L. and Cai, T., 2022. [Efficient Evaluation of Prediction Rules in Semi-Supervised Settings under Stratified Sampling](#). Journal of the Royal Statistical Society: Series B, Statistical Methodology, 84(4), pp.1353-1391.

## Semi-supervised model evaluation: Problem setting

- Observable Data
  - ★ Labeled:  $\mathcal{L} = \{(Y_i, \mathbf{X}_i^T)^\top \mid i = 1, \dots, n\}$
  - ★ Unlabeled:  $\mathcal{U} = \{\mathbf{X}_i \mid i = n + 1, \dots, n + N\}$
- Assumptions
  - ★ Labeling is independent of  $Y$  and  $\mathbf{X}$  (ie. MCAR)
  - ★  $n/N \rightarrow 0$  as  $n \rightarrow \infty$

## Semi-supervised model evaluation: Problem setting

- Goal

Evaluate the classification rule for a future  $Y^0$  given by

$$\hat{Y} = I(\mathcal{P}_{\hat{\theta}}^0 > c)$$

where  $\mathcal{P}_{\hat{\theta}}^0 = g(\hat{\theta}^\top \vec{X}^0)$  is from fitting a *working* GLM

$$P(Y = 1 \mid \mathbf{X}) = g(\alpha + \beta^\top \mathbf{X}) = g(\boldsymbol{\theta}^\top \vec{\mathbf{X}})$$

- Focus on estimation of the ROC parameters

eg.  $\overline{\text{TPR}}(c) = P(\mathcal{P}_{\hat{\theta}}^0 > c \mid Y^0 = 1)$

## (Some) Related work

### Missing data

Rubin (1976), Fluss et al (2009), Rotnitzky et al (2011),  
Zawistowski et al (2017), Tan et al (2019)

- In our setting,  $n/N \rightarrow 0$  as  $n \rightarrow \infty$ 
  - ★ The distribution of  $\mathbf{X}$  is known due to size of  $\mathcal{U}$
  - ★ Existing methods rely on the positivity assumption

## (Some) Related work

### Semi-supervised learning

Cozman et al (2003), Wasserman et al (2007),  
Sokolovska et al (2008), Chakrabortty et al (2020)

- Focus is on model estimation when the outcome is MCAR
  - ★ Our goal is to make model evaluation more precise

# Why focus on model evaluation?

---

## “Life after modeling”

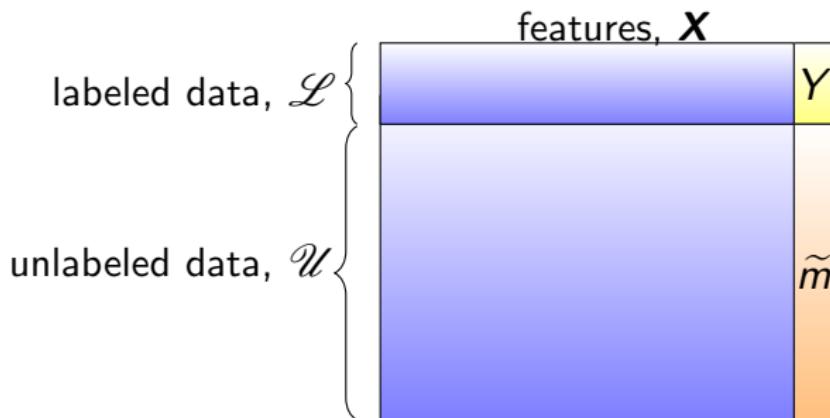
- Model errors impact downstream use
- Model evaluation requires sufficient labeled data to yield precise (ie. low variance) estimation
- Without precise evaluation:
  - ★ Model errors can be misunderstood
  - ★ Differences across subpopulations can go undetected
  - ★ Model comparisons can be unreliable

## Overview: Semi-supervised model evaluation

Goal Enable precise model evaluation with a small  $\mathcal{L}$

Observation  $\mathcal{U}$  can be incorporated into estimation

Approach Use  $\mathcal{L}$  to impute the missing outcomes with  $\tilde{m}$



## Motivating our approach: A simple example

Consider estimating  $\mu = E(Y)$  with information on a single  $X$ .

- The familiar (supervised) estimator of  $\mu$  is

$$\hat{\mu}_{\text{SL}} = n^{-1} \sum_{i=1}^n Y_i$$

- We can make use of  $\mathcal{U}$  by noting that

$$\mu = E(Y) = E\{E(Y | X)\} = \int m(x) dF(x)$$

## Motivating our approach: A simple example

- Proposal: Take the empirical counterpart of  $\int m(x)dF(x)$

$$\hat{\mu}_{ss} = N^{-1} \sum_{i=n+1}^{n+N} \tilde{m}(X_i) \text{ where } \tilde{m}(x) = \frac{\sum_{j=1}^n K_h(X_j - x) Y_j}{\sum_{j=1}^n K_h(X_j - x)}$$

and  $\tilde{m}(\cdot)$  is the Nadaraya-Watson estimator

## Motivating our approach: A simple example

- Proposal: Take the empirical counterpart of  $\int m(x)dF(x)$

$$\hat{\mu}_{\text{ss}} = N^{-1} \sum_{i=n+1}^{n+N} \tilde{m}(X_i) \text{ where } \tilde{m}(x) = \frac{\sum_{j=1}^n K_h(X_j - x) Y_j}{\sum_{j=1}^n K_h(X_j - x)}$$

and  $\tilde{m}(\cdot)$  is the Nadaraya-Watson estimator

- Result: We can show that

$$\sqrt{n}(\hat{\mu}_{\text{SL}} - \mu) = n^{-1/2} \sum_{i=1}^n (Y_i - \mu) + o_p(1)$$

while

$$\sqrt{n}(\hat{\mu}_{\text{ss}} - \mu) = n^{-1/2} \sum_{i=1}^n \{Y_i - m(X_i)\} + o_p(1)$$

## Motivating our approach: A simple example

- This implies

$$\text{Var}(\hat{\mu}_{\text{SL}}) = \text{Var}(Y) = E\{\text{Var}(Y | X)\} + \text{Var}\{E(Y | X)\}$$

and

$$\text{Var}(\hat{\mu}_{\text{ss}}) = E\{\text{Var}(Y | X)\}$$

The semi-supervised estimator is asymptotically  
more efficient than the supervised estimator  
when  $\text{Var}\{E(Y | X)\} > 0$

## Our imputation-based approach to SS learning

With more complex parameters we aim to balance:

- Flexibility: Impute  $Y$  with a method that captures the dependency of  $Y$  on  $X$  to enhance our ability to gain **efficiency**
- Feasibility: Impute  $Y$  with a method that is **robust** to potential misspecification of the imputation model

## Motivating the SS estimator of the TPR

- Recall

$$\overline{\text{TPR}}(c) = P(\mathcal{P}_{\hat{\theta}}^0 > c \mid Y^0 = 1) = \frac{E\{Y^0 I(\mathcal{P}_{\hat{\theta}}^0 > c)\}}{E(Y^0)}$$

- The supervised estimator of  $\overline{\text{TPR}}(c)$  is

$$\widehat{\text{TPR}}_{\text{SL}}(c) = \frac{\sum_{i=1}^n I(\mathcal{P}_{\hat{\theta}i} > c) Y_i}{\sum_{i=1}^n Y_i}$$

## Motivating the SS estimator of the TPR

- Similar to the estimation of  $\mu$ ,

$$\overline{\text{TPR}}(c) = \frac{\mathbb{E}\{Y^0 I(\mathcal{P}_{\hat{\theta}}^0 > c)\}}{\mathbb{E}(Y^0)} = \frac{\mathbb{E}\{\bar{m}(\mathcal{P}_{\hat{\theta}}^0) I(\mathcal{P}_{\hat{\theta}}^0 > c)\}}{\mathbb{E}\{\bar{m}(\mathcal{P}_{\hat{\theta}}^0)\}}$$

where  $\bar{m}(s) = P(Y^0 = 1 | \mathcal{P}_{\hat{\theta}}^0 = s)$

## Motivating the SS estimator of the TPR

- Similar to the estimation of  $\mu$ ,

$$\overline{\text{TPR}}(c) = \frac{\mathbb{E}\{Y^0 I(\mathcal{P}_{\hat{\theta}}^0 > c)\}}{\mathbb{E}(Y^0)} = \frac{\mathbb{E}\{\bar{m}(\mathcal{P}_{\hat{\theta}}^0) I(\mathcal{P}_{\hat{\theta}}^0 > c)\}}{\mathbb{E}\{\bar{m}(\mathcal{P}_{\hat{\theta}}^0)\}}$$

where  $\bar{m}(s) = P(Y^0 = 1 | \mathcal{P}_{\hat{\theta}}^0 = s)$

1.  $\overline{\text{TPR}}(c)$  depends on the distribution of  $\mathbf{X}$
2.  $Y$  may be imputed using  $\bar{m}(\cdot)$  estimated from  $\mathcal{L}$

## Semi-supervised estimation of the TPR (ssROC)

1. Estimate  $\bar{m}(s) = P(Y^0 = 1 | \mathcal{P}_{\hat{\theta}} = s)$  with  $\mathcal{L}$  as

$$\tilde{m}(s) = \frac{\sum_{j=1}^n K_h(\mathcal{P}_{\hat{\theta}j} - s) Y_j}{\sum_{j=1}^n K_h(\mathcal{P}_{\hat{\theta}j} - s)}$$

where  $K_h(u) = h^{-1}K(u/h)$ ,  $K(\cdot)$  is a smooth symmetric kernel function, and  $h$  is the bandwidth with  $nh^2 \rightarrow$  and  $nh^4 \rightarrow \infty$  as  $n \rightarrow 0$  and

2. Estimate  $\widehat{\text{TPR}}(c)$  with  $\mathcal{U}$  as

$$\widehat{\text{TPR}}_{\text{ss}}(c) = \frac{\sum_{i=n+1}^{n+N} I(\mathcal{P}_{\hat{\theta}i} > c) \tilde{m}(\mathcal{P}_{\hat{\theta}i})}{\sum_{i=n+1}^{n+N} \tilde{m}(\mathcal{P}_{\hat{\theta}i})}$$

## Justification for ssROC

Under standard regularity conditions and under-smoothing,  $\sqrt{n}\{\widehat{\text{ROC}}_{\text{ss}}(u_0) - \overline{\text{ROC}}(u_0)\}$  is equivalent to

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathcal{G}_{u_0}(\mathcal{P}_{\theta_0 i}) \{Y_i - E(Y_i | \mathcal{P}_{\theta_0 i})\} - \mathcal{J}_{u_0}(D_i) + o_p(1)$$

while  $\sqrt{n}\{\widehat{\text{ROC}}_{\text{SL}}(u_0) - \overline{\text{ROC}}(u_0)\}$  is equivalent to

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathcal{G}_{u_0}(\mathcal{P}_{\theta_0 i}) \{Y_i - \mu\} - \mathcal{J}_{u_0}(D_i) + o_p(1)$$

## Justification for ssROC

Under standard regularity conditions and under-smoothing,  $\sqrt{n}\{\widehat{\text{ROC}}_{\text{ss}}(u_0) - \overline{\text{ROC}}(u_0)\}$  is equivalent to

$$n^{-\frac{1}{2}} \sum_{i=1}^n \underbrace{\mathcal{G}_{u_0}(\mathcal{P}_{\theta_0 i})\{Y_i - E(Y_i | \mathcal{P}_{\theta_0 i})\}}_{\text{Accuracy Measure}} - \underbrace{\mathcal{J}_{u_0}(D_i)}_{\hat{\theta}} + o_p(1)$$

while  $\sqrt{n}\{\widehat{\text{ROC}}_{\text{SL}}(u_0) - \overline{\text{ROC}}(u_0)\}$  is equivalent to

$$n^{-\frac{1}{2}} \sum_{i=1}^n \mathcal{G}_{u_0}(\mathcal{P}_{\theta_0 i})\{Y_i - \mu\} - \mathcal{J}_{u_0}(D_i) + o_p(1)$$

## ssROC in action: Data analysis overview

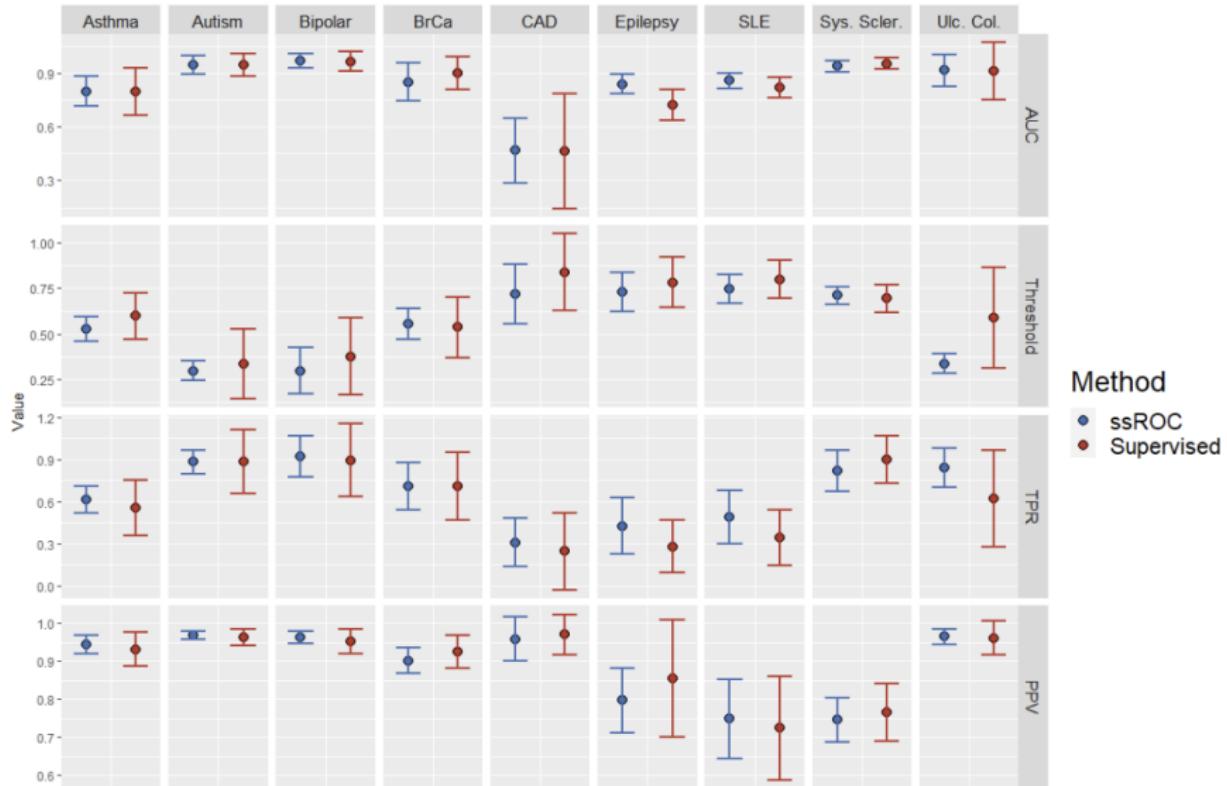
- Evaluated models for 9 phenotypes from Mass General Brigham Biobank
- Models trained with PheNorm algorithm
- Compared supervised ROC analysis and ssROC
- Evaluated predictive performance based on the TPR, PPV, and threshold with FPR = 0.10 and the AUC
- Used perturbation resampling for variance estimation

Note: These are preliminary results.

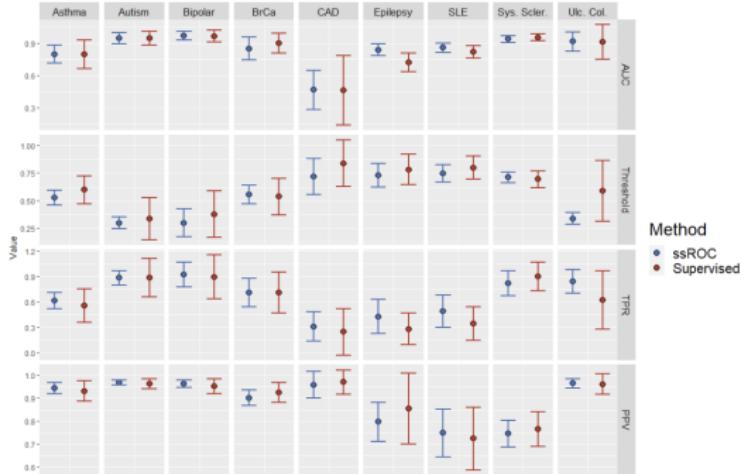
# MGB Phenotypes

Phenotype	<i>n</i>	<i>N</i>	<i>P</i>
Asthma	201	313816	0.39
Autism	170	18955	0.68
Bipolar	110	65052	0.35
Breast Cancer	110	102953	0.57
Coronary Atherosclerosis	157	202293	0.41
Epilepsy	87	47233	0.56
Systemic Lupus Erythematosus	97	15111	0.34
Systemic Sclerosis	189	4083	0.43
Ulcerative Colitis	132	27351	0.42

# Analysis results

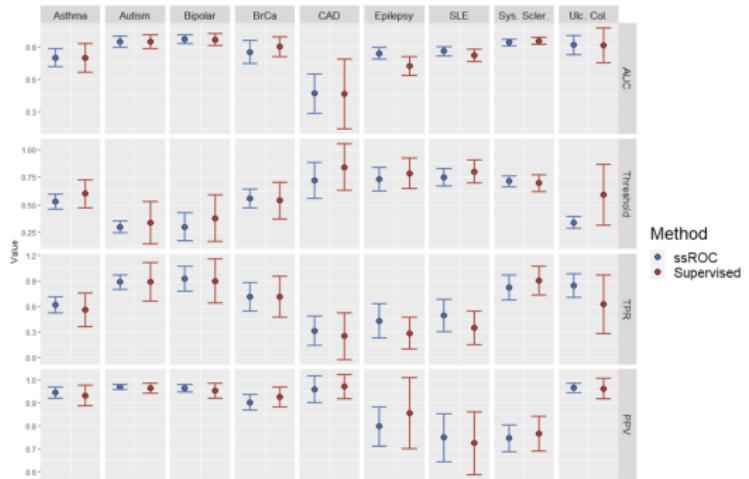


# Analysis results



On average, confidence intervals from ssROC are half the length from supervised ROC analysis.

# Analysis results



A confidence interval with the same length of supervised ROC analysis can be obtained with **a third of the labeled data** using ssROC.

## Summary

---

- EHR data is a valuable resource for clinical research
- Phenotyping is a fundamental aspect of EHR research, but is bottlenecked by labeled data constraints
- ssROC enables precise model evaluation with limited labeled data
- Future directions
  - ★ Non-random sampling of  $\mathcal{L}$  (eg. transfer learning)
  - ★ Evaluation of fairness gaps

# Thank you!

---

You can find my slides at

[https://github.com/jlgrons/CMS-Winter-Meeting-2022.](https://github.com/jlgrons/CMS-Winter-Meeting-2022)