Leveraging electronic health records for data science

Jesse Gronsbell
Department of Statistical Sciences
University of Toronto

CMS Winter Meeting 2022

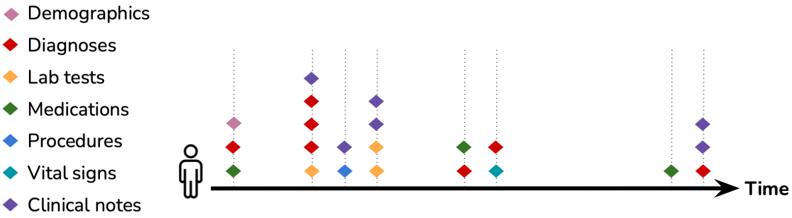
Stochastic Systems, Probability, and Other Mathematical Aspects of Data Science

Roadmap for today

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

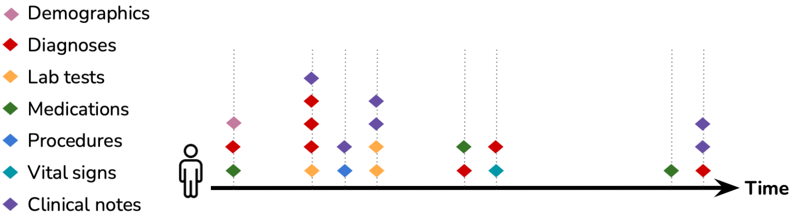
What is an Electronic Health Record (EHR)?

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



What is an Electronic Health Record (EHR)?

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



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
$\uparrow$ **Available**	Increasing EHR adoption worldwide

The opportunity: Learn from EHR data

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

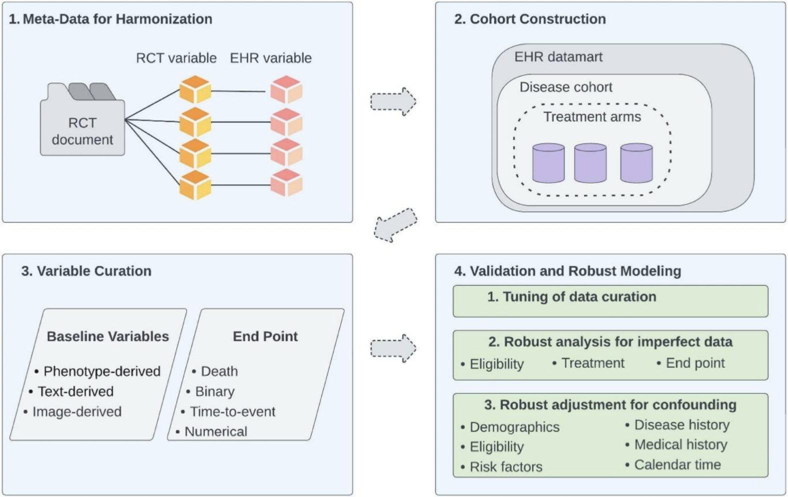
Data science at bedside: A “Green Button”

> “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

Stanford Green Button Project

Data science at the bench: Real-world evidence



Hou J, Zhao R, Gronsbell J, et al. *Harnessing electronic health records for real-world evidence.* arXiv:2211.16609

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, etc. . . .

Phenotypes are the foundation of EHR research

Review

OPEN ACCESS

A review of approaches to identifying patient phenotype cohorts using electronic health records

Chaitanya Shivade,[1] Preethi Raghavan,[1] Eric Fosler-Lussier,[1] Peter J Embi,[2] Noemie Elhadad,[3] Stephen B Johnson,[4] Albert M Lai[2]

Annual Review of Biomedical Data Science

Advances in Electronic Phenotyping: From Rule-Based Definitions to Machine Learning Models

Juan M. Banda,* Martin Seneviratne,*
Tina Hernandez-Boussard, and Nigam H. Shah

Phenotypes are the foundation of EHR research

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

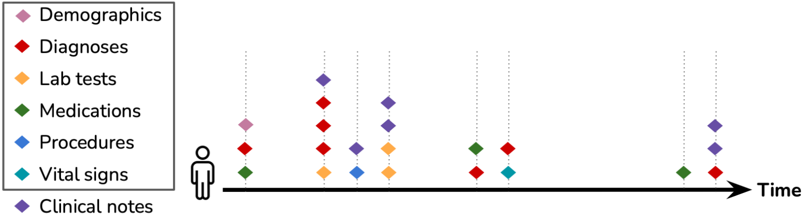
Siyue Yang[1], Paul Varghese[2], Ellen Stephenson[3], Karen Tu[3], and Jessica Gronsbell[1,3,4]

Why is phenotyping challenging?

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

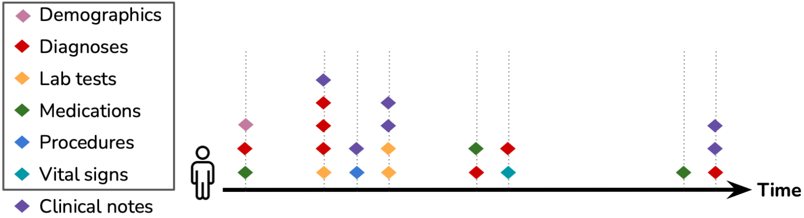
Leo Anthony Celi

The two flavors of EHR data



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

The two flavors of EHR data

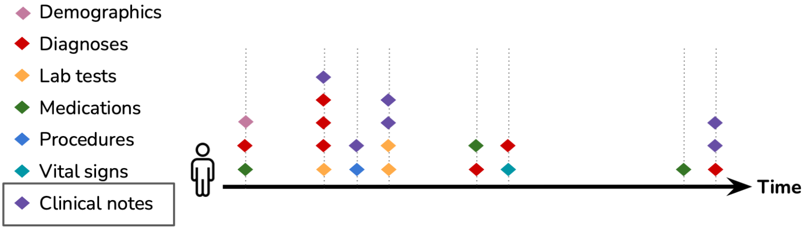


1. **Structured data**: Easy to extract, but lacks context

eg. ICD diagnosis code $\not\rightarrow$ disease diagnosis

complex diagnosis, upcoding, temporal shift, etc.

The two flavors of EHR data

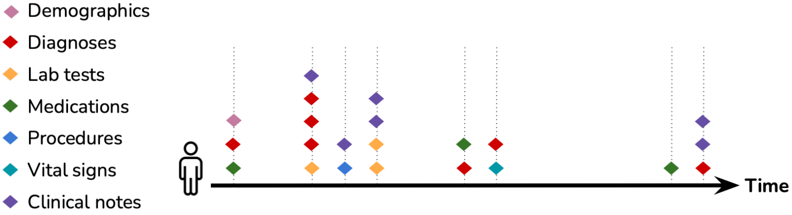


2. Unstructured data: Rich information, but requires NLP

clinical terms $\rightarrow$ concept unique identifier (CUI)

eg. "Rheumatoid Arthritis", "RA" $=$ C000387

The two flavors of EHR data



$$\boxed{\text{phenotype} \approx \text{structured} + \text{unstructured data}}$$

Solution: ML-based phenotyping algorithms

1. Select relevant features, $\boldsymbol{X}$
2. Manually review records to label phenotype status, Y
3. Build the model, $Y \sim g(\boldsymbol{X}; \boldsymbol{\theta}) \rightarrow \hat{Y}(\boldsymbol{X}) \sim g(\boldsymbol{X}; \hat{\boldsymbol{\theta}})$

Solution: ML-based phenotyping algorithms

1. Select relevant features, $\boldsymbol{X}$
2. Manually review records to label phenotype status, Y
3. Build the model, $Y \sim g(\boldsymbol{X}; \boldsymbol{\theta}) \rightarrow \hat{Y}(\boldsymbol{X}) \sim g(\boldsymbol{X}; \hat{\boldsymbol{\theta}})$

1^{st} algorithm for rheumatoid arthritis

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

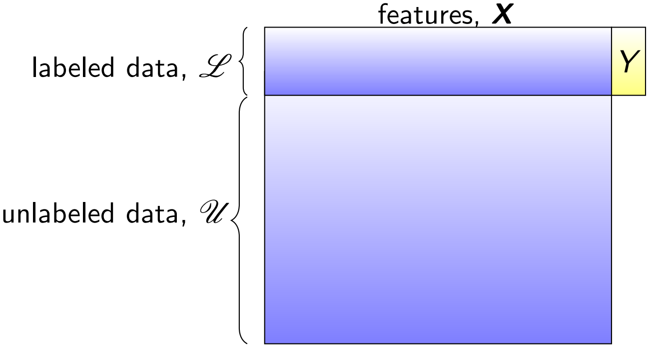
Solution: ML-based phenotyping algorithms

1. Select relevant features, $\boldsymbol{X}$
2. Manually review records to label phenotype status, Y
3. Build the model, $Y \sim g(\boldsymbol{X}; \boldsymbol{\theta}) \rightarrow \hat{Y}(\boldsymbol{X}) \sim g(\boldsymbol{X}; \hat{\boldsymbol{\theta}})$

"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. <u>Model estimation</u> (eg. regression model)
2. Predictive performance evaluation (eg. ROC parameters)

How can we make phenotyping more efficient?

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

1. <u>Model estimation</u> (eg. regression model)
2. Model evaluation (eg. ROC parameters)

ORIGINAL ARTICLE

WILEY Biometrics

Automated feature selection of predictors in electronic medical records data

Jessica Gronsbell[1] | **Jessica Minnier**[2,*] | **Sheng Yu**[3] | **Katherine Liao**[4] | **Tianxi Cai**[5]

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)

Journal of the American Medical Informatics Association, 25(1), 2018, 54–60
doi: 10.1093/jamia/ocx111
Advance Access Publication Date: 3 November 2017
Research and Applications

Research and Applications

Enabling phenotypic big data with PheNorm

Sheng Yu,[1,2] Yumeng Ma,[3] Jessica Gronsbell,[4] Tianrun Cai,[5] Ashwin N Ananthakrishnan,[6] Vivian S Gainer,[7] Susanne E Churchill,[8] Peter Szolovits,[9] Shawn N Murphy,[7,10] Isaac S Kohane,[8] Katherine P Liao,[11] and Tianxi Cai[4]

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)

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 3, *pp.* 579–594

Semi-supervised approaches to efficient evaluation of model prediction performance

Jessica L. Gronsbell and Tianxi Cai

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)

ORIGINAL ARTICLE

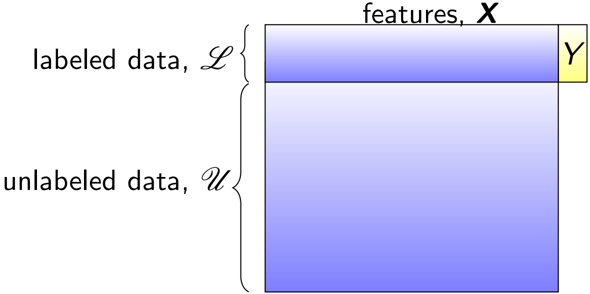
Efficient evaluation of prediction rules in semi-supervised settings under stratified sampling

Jessica Gronsbell[1] | Molei Liu[2] | Lu Tian[3] | Tianxi Cai[2,4]

Overview: Semi-supervised model evaluation

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

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



Approach Use $\mathscr{L}$ to impute the missing outcomes, Y

Problem setting

- Observable Data
  - ⋆ Labeled: $\mathscr{L} = \{(Y_i, \boldsymbol{X}_i^\mathsf{T})^\mathsf{T} \mid i = 1, \ldots, n\}$
  - ⋆ Unlabeled: $\mathscr{U} = \{\boldsymbol{X}_i \mid i = n+1, \ldots, n+N\}$

- Assumptions
  - ⋆ $\mathscr{L} \perp \mathscr{U}$
  - ⋆ Labeling is independent of Y and $\boldsymbol{X}$ (ie./ MCAR)
  - ⋆ $n/N \to 0$ as $n \to \infty$

Problem setting

- <u>Goal</u>

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

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

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

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

- Focus on estimation of the ROC parameters

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

(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, $n/N \to 0$ as $n \to \infty$
  - ⋆ The distribution of $\boldsymbol{X}$ is known due to size of $\mathscr{U}$
  - ⋆ 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
- Models aren’t the only part of the ML pipeline that are labeled data hungry
- Without precise evaluation:
  - ⋆ Model errors can be misunderstood
  - ⋆ Differences across subpopulations can go undetected
  - ⋆ Model comparisons are unreliable

Why it makes sense: A simple example

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

- The familiar (supervised) estimator of μ is

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

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

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

Why it makes sense: A simple example

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

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

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

Why it makes sense: A simple example

- <u>Proposal</u>: 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

- <u>Result</u>: 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

$$\sqrt{n}(\hat{\mu}_{\text{SS}} - \mu) = n^{-1/2} \sum_{i=1}^{n} \{y_i - m(x_i)\} + o_p(1)$$

Why it makes sense: A simple example

- This implies

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

and

$$\text{Var}(\hat{\mu}_{\text{SS}}) = \text{E}\{\text{Var}(Y \mid X)\}$$

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

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 $\boldsymbol{X}$ 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

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

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

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

Motivating the SS estimator of the TPR

- Similar to the estimation of μ,

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

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

Motivating the SS estimator of the TPR

- Similar to the estimation of μ,

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

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

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

Semi-supervised estimation of the TPR (ssROC)

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

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

   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 $\overline{\mathrm{TPR}}(c)$ with $\mathscr{U}$ as

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

Justification for ssROC

Under standard regularity conditions and undersmoothing, $\sqrt{n}\{\widehat{\mathrm{ROC}}_{\mathrm{SS}}(u_0) - \overline{\mathrm{ROC}}(u_0)\}$ is equivalent to

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{G}_{u_0}(\mathcal{P}_{\boldsymbol{\theta}_0 i})\{y_i - E(y_i \mid \mathcal{P}_{\boldsymbol{\theta}_0 i})\} - \mathcal{J}_{u_0}(\boldsymbol{D}_i) + o_p(1)$$

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

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{G}_{u_0}(\mathcal{P}_{\boldsymbol{\theta}_0 i})\{y_i - \mu\} - \mathcal{J}_{u_0}(\boldsymbol{D}_i) + o_p(1)$$

Justification for ssROC

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

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \overbrace{\mathcal{G}_{u_0}(\mathcal{P}_{\boldsymbol{\theta}_0 i})\{y_i - E(y_i \mid \mathcal{P}_{\boldsymbol{\theta}_0 i})\}}^{\text{Accuracy Measure}} - \overbrace{\mathcal{J}_{u_0}(\boldsymbol{D}_i)}^{\hat{\boldsymbol{\theta}}} + o_p(1)$$

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

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} \mathcal{G}_{u_0}(\mathcal{P}_{\boldsymbol{\theta}_0 i})\{y_i - \mu\} - \mathcal{J}_{u_0}(\boldsymbol{D}_i) + o_p(1)$$

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 NPV with FPR = 0.10 and the AUC
- Used perturbation resampling for variance estimation

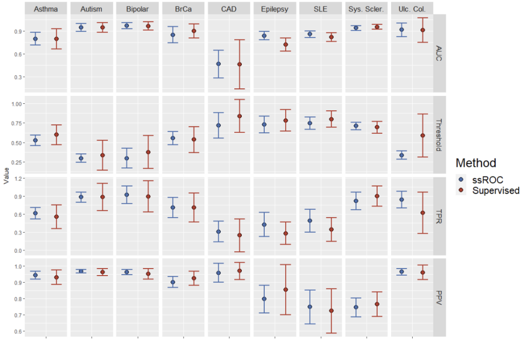
Note: These are preliminary results.

Hong C, Bonzel CL, Gao J, Varghese P, Zakir K, Deng J, and Gronsbell J. *ssROC: Semi-Supervised ROC Analysis for Reliable and Streamlined Evaluation of Phenotyping Algorithms.* Preprint forthcoming.

MGB Phenotypes

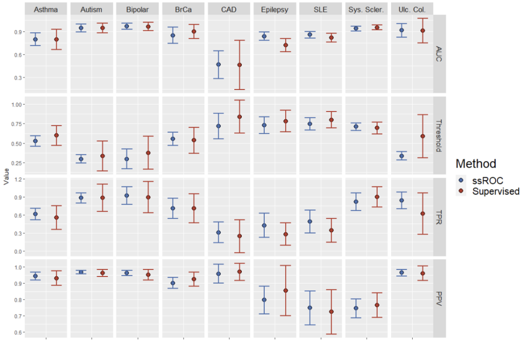
Phenotype	n	N	P
Asthma	201	313816	0.39
Autism	170	18955	0.68
Bipolar	110	65052	0.35
Breast Cancer	110	102953	0.57
Coronary Atherosclerosis	157	202293	0.41
Epilepsy	87	47233	0.56
Systemic Lupus Erythematosus	97	15111	0.34
Systemic Sclerosis	189	4083	0.43
Ulcerative Colitis	132	27351	0.42

Results



On average, confidence intervals from ssROC are half the length from supervised ROC 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 $\mathscr{L}$ (eg. transfer learning)
  - Evaluation of fairness gaps
  - Model comparisons
  - Using a surrogate for Y (eg. medical imaging)
  - Parametric estimation for small $\mathscr{L}$

Future directions

- Non-random sampling of $\mathscr{L}$ (eg. transfer learning)
- Evaluation of fairness gaps
- Model comparisons
- Incorporating a surrogate for Y (eg. medical imaging)
- Parametric estimation for small $\mathscr{L}$

Thank you!

You can find my slides at
https://github.com/jlgrons/CMS-Winter-Meeting-2022.