Statistical learning with high volume, high noise health data

Jesse Gronsbell

Department of Statistical Sciences
University of Toronto

WAM 2022, Institute for Advanced Study
May 27, 2022

Today's roadmap

- Who I am
- How I got here
- What I work on
- Why my work matters

Who I am

Assistant Professor in Statistics, University of Toronto

Who I am

Assistant Professor in Statistics, University of Toronto

I develop statistical learning methods for
high volume, high noise health data

But first, how I got here

A very long time ago

I didn't think I would go to college

How I got here

Fired from my job as a line chef at El Azteco

How I got here

A very long time ago

I didn't think I would go to college

Berkeley, BA in Applied Mathematics

Harvard, PhD in Biostatistics with Tianxi Cai

Stanford, Postdoc in Biomedical Data Science with Lu Tian

Alphabet's Verily Life Sciences, Data Scientist

University of Toronto, Assistant Professor in Statistics

Now

How I figured it out

I took every opportunity in statistics that was given to me

How I figured it out

I took every opportunity in statistics that was given to me

- Baseball pitch classification
- Ballistic missile defense
- Electronic health records
- Mobile health
- COVID-19 testing strategies
- ...

What I learned: Part I

It's easier to take risks when you let your inhibitions go

What I learned: Part II

"The best thing about being a statistician is that you get to play in everyone's backyard."

John Tukey

Where I'm playing now

I develop statistical learning methods for high volume, high noise health data

Where I'm playing now

I develop statistical learning methods for high volume, high noise health data

Where I'm playing now

I develop statistical learning methods for
high volume, high noise health data

- Machine learning (ML)
  Development of models & algorithms from data
- Statistical learning
  Branch of applied statistics that emerged in response to ML focusing on statistical models & uncertainty quantification

Very similar, but with different emphases.

Where I'm playing now

I develop statistical learning methods for
high volume, high noise health data

Where I'm playing now

I develop statistical learning methods for
high volume, high noise health data

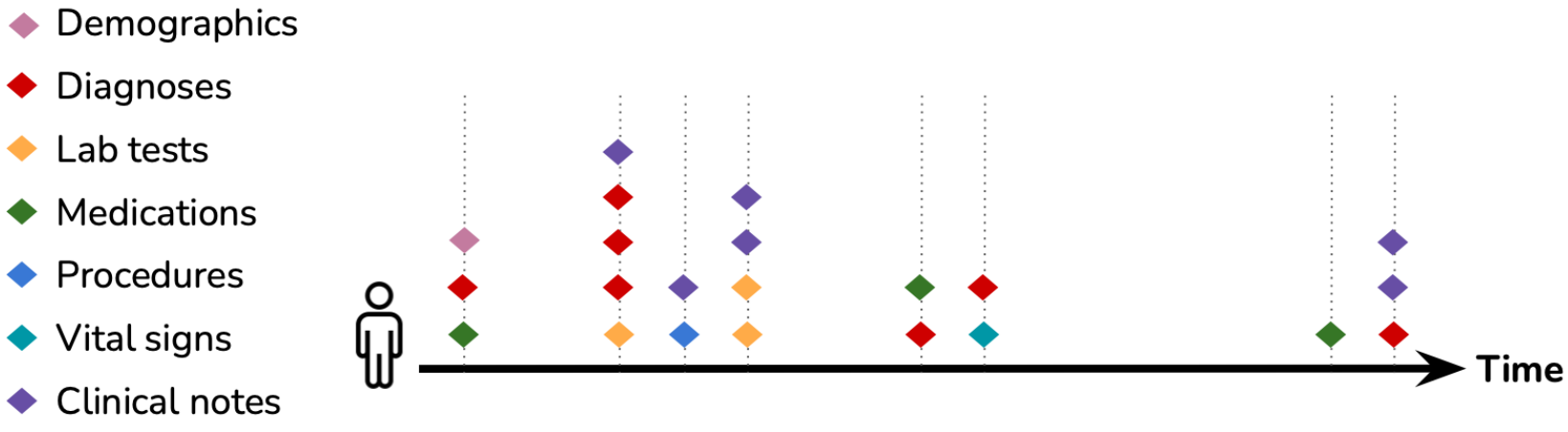
Electronic health record (EHR) data

Where I'm playing now

I develop statistical learning methods for
high volume, high noise health data

Electronic health record (EHR) data

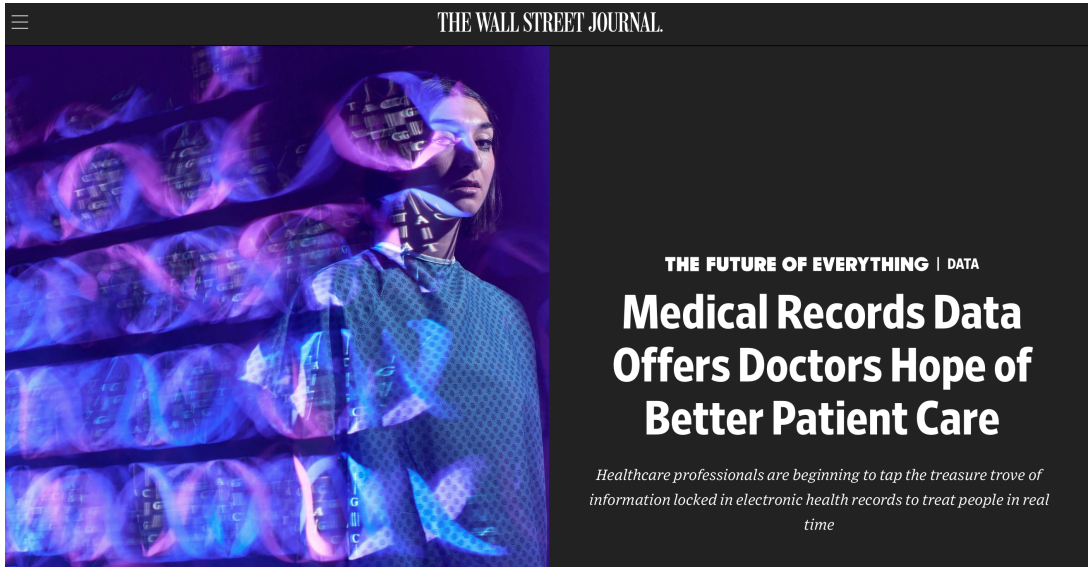
An EHR is an electronic record of a
patient's interactions with a healthcare system



EHR data is high in volume

Big	Longitudinal records on large populations
Detailed	Information on numerous fields
Representative	Real-world patients
↑ **Available**	Increasing EHR adoption worldwide

The dream: Use EHR data to provide better care



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

A "Green Button" for patients like mine

"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

A "Green Button" for patients like mine

"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

This Green Button is still an aspiration

EHR data is high in noise

Finding "patients like mine" requires
identifying patients with a similar phenotype

Phenotype: patient characteristic inferred from their EHR

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

EHR data is high in noise

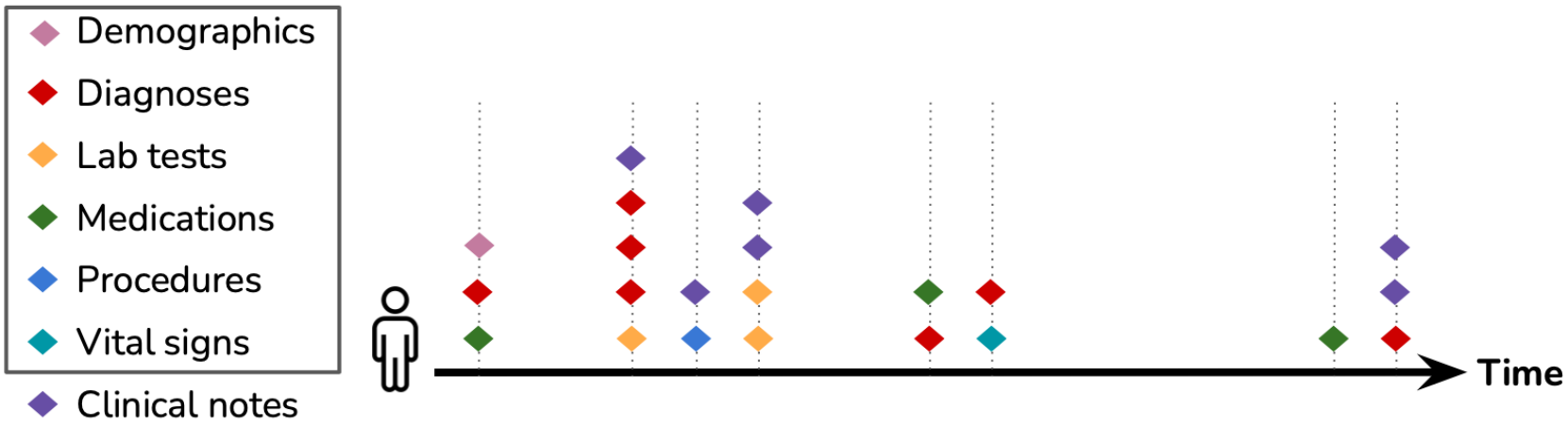
Finding "patients like mine" requires
identifying patients with a similar phenotype

Phenotype: patient characteristic inferred from their EHR

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

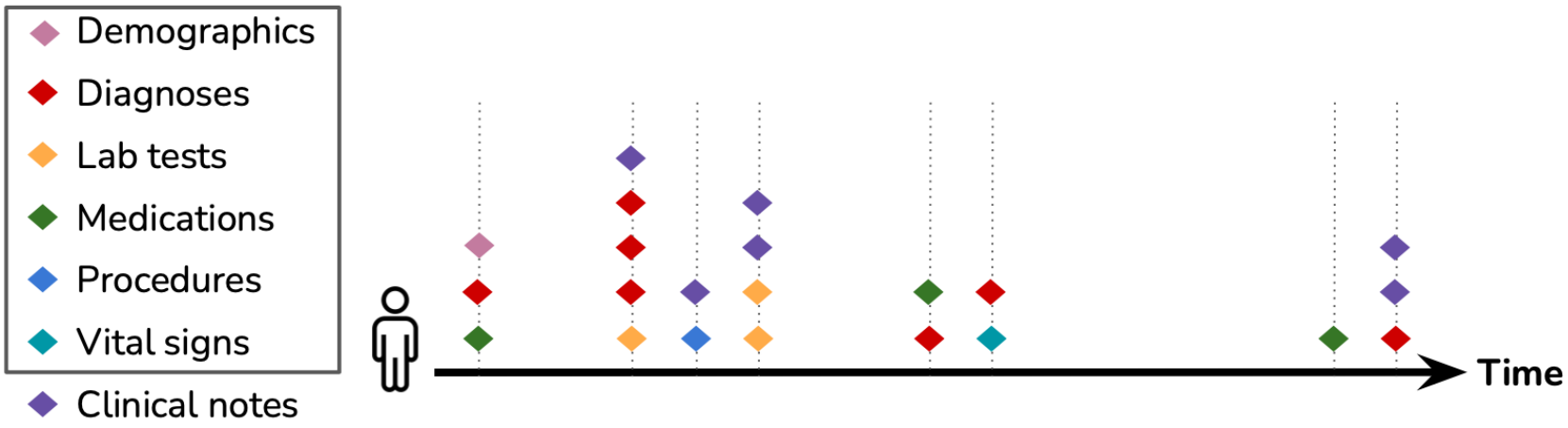
EHRs do not have explicit information on phenotypes

The two flavors of EHR data



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

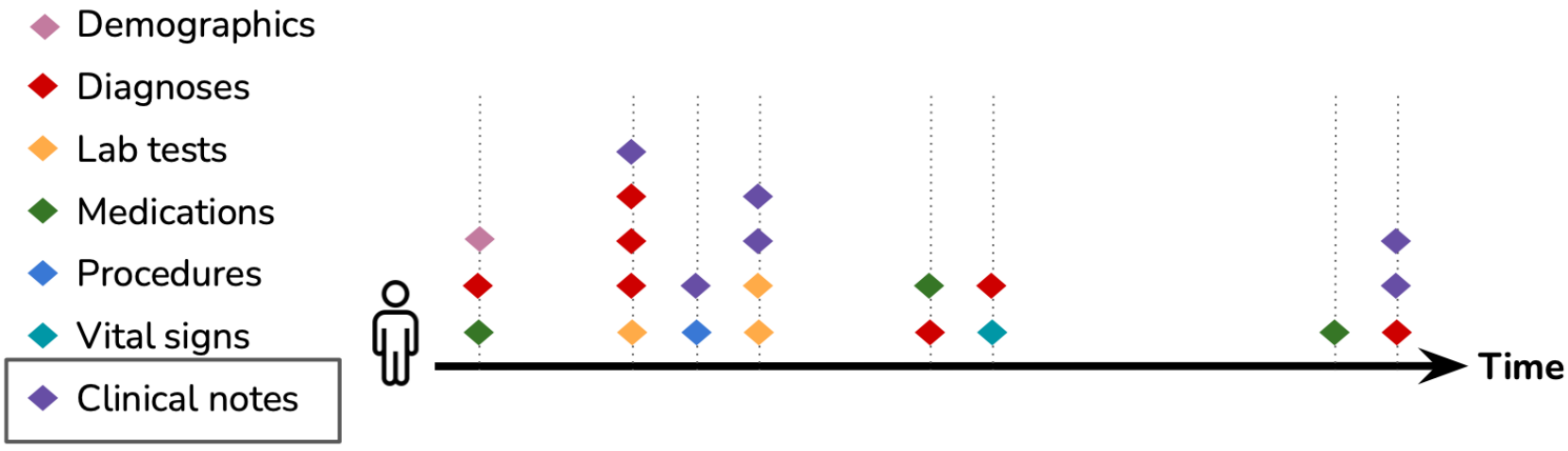
The two flavors of EHR data



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

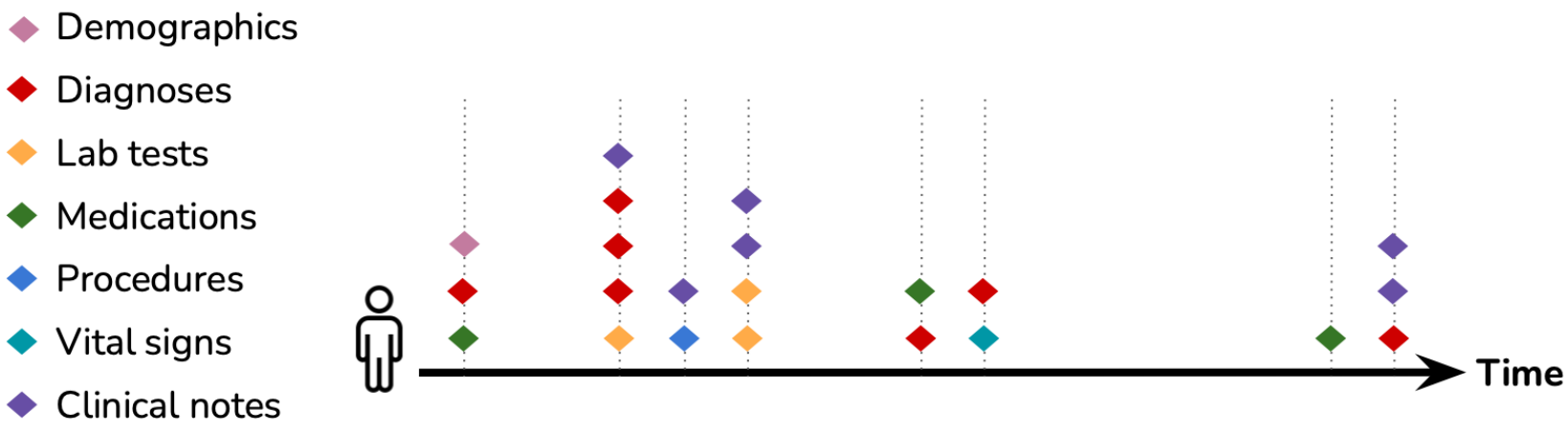
   eg. ICD diagnosis code: Upcoding

The two flavors of EHR data



2. **Unstructured data**: Detailed, but requires NLP

The two flavors of EHR data



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

Potential solution: Statistical/ML models

Derive **features** from EHRs

$+$

Manually review EHRs to obtain the **label**

$\downarrow$

Train a supervised statistical/ML model to infer the phenotype

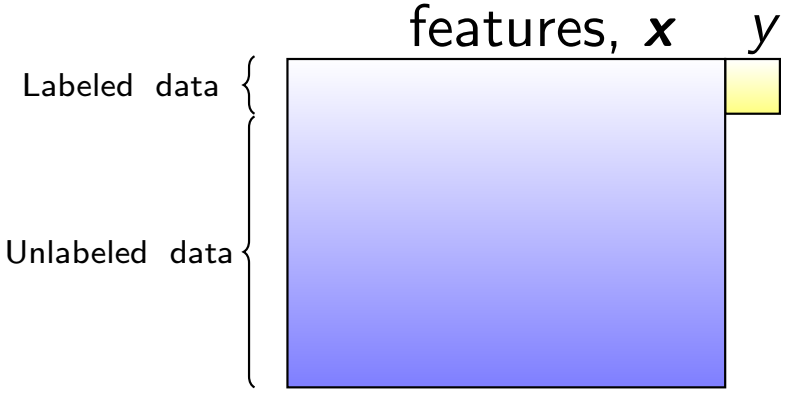
Challenge: Obtaining labeled data

Labeling health records is the worst thing I have done at Alphabet.

- A physician colleague

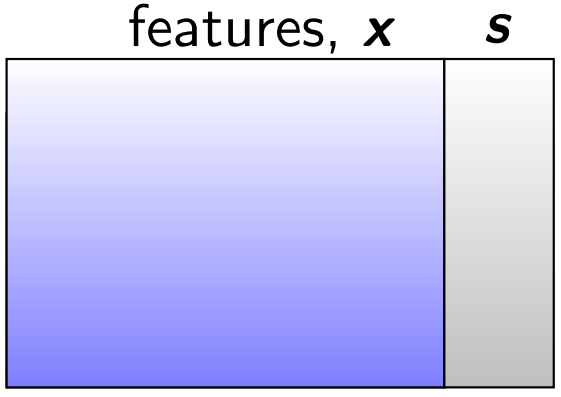
My work: Learning with limited/noisy labeled data

1. Semi-supervised learning



Utilize labeled and unlabeled data to improve statistical learning

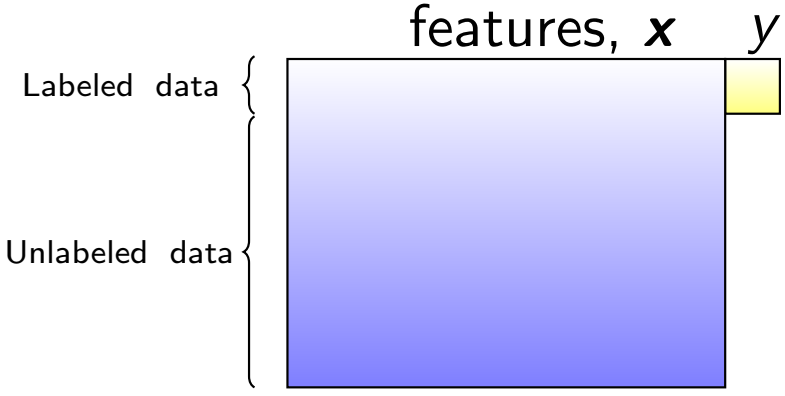
2. Weakly-supervised learning



Utilize noisy labels automatically extracted from EHRs to guide learning in place of true labels

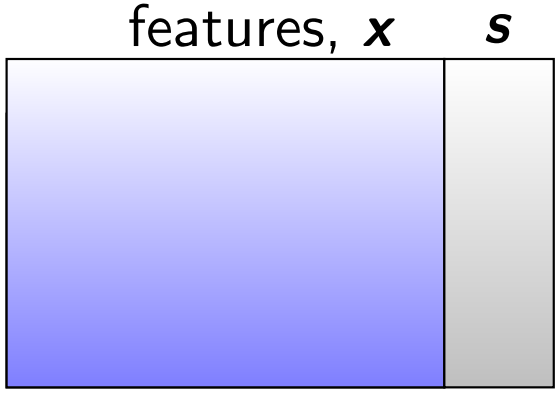
My work: Learning with limited/noisy labeled data

1. Semi-supervised learning



Utilize labeled and unlabeled data to improve statistical learning

2. Weakly-supervised learning



Utilize noisy labels automatically extracted from EHRs to guide learning in place of true labels

Enable accurate statistical learning with very little labeled data

Why my work matters

We can't use EHR data in an impactful way without accurate phenotypes that are easy to obtain

Example: EHR-based registry for rare diseases

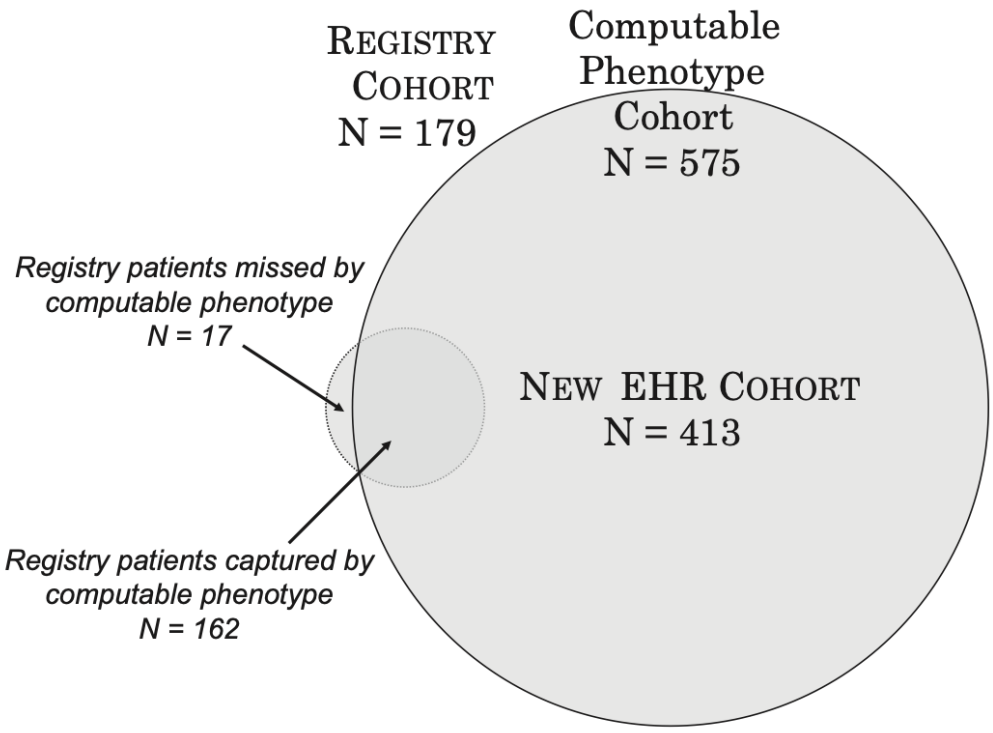
ORIGINAL ARTICLES

www.jpeds.com • THE JOURNAL OF PEDIATRICS

A Computable Phenotype Improves Cohort Ascertainment in a Pediatric Pulmonary Hypertension Registry

Alon Geva, MD, MPH[1,2,3], Jessica L. Gronsbell, BA[4], Tianxi Cai, ScD[4], Tianrun Cai, MD[5], Shawn N. Murphy, MD, PhD[6,7,8], Jessica C. Lyons, MS[8], Michelle M. Heinz, BS[1], Marc D. Natter, MD[1,9], Nandan Patibandla, MS[10], Jonathan Bickel, MD, MS[1,9,10], Mary P. Mullen, MD, PhD[9,11], and Kenneth D. Mandl, MD, MPH[1,8,9], for the Pediatric Pulmonary Hypertension Network and National Heart, Lung, and Blood Institute Pediatric Pulmonary Vascular Disease Outcomes Bioinformatics Clinical Coordinating Center Investigators*

Example: EHR-based registry for rare diseases



Identify large cohorts of patients for clinical studies in a fraction of the time and cost

None of this is possible without a team!

- Tianxi Cai (Harvard)
- Lu Tian (Stanford)
- Paul Varghese (Verily)
- Katherine Liao (Brigham and Women's Hospital)
- Chuan Hong (Duke)
- Molei Liu (Harvard)
- Jessica Minnier (OHSU)

Thanks!

j.gronsbell@utoronto.ca

Relevant References

Semi-supervised learning

- Gronsbell J, Liu M, Tian L, Cai T. Efficient Estimation and Evaluation of Prediction Rules in Semi-Supervised Settings under Stratified Sampling. Accepted at Journal of the Royal Statistical Society: Series B (Statistical Methodology). [link]
- Gronsbell J, Cai T. Semi-supervised approaches to efficient evaluation of model prediction performance. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2018 Jun;80(3):579-94. [link]

Relevant References

Weakly-supervised learning

- Gronsbell J, Minnier J, Yu S, Liao K, Cai T. Automated feature selection of predictors in electronic medical records data. Biometrics. 2019 Mar;75(1):268-77. [link]
- Yu S, Ma Y, Gronsbell J, Cai T, Ananthakrishnan AN, Gainer VS, Churchill SE, Szolovits P, Murphy SN, Kohane IS, Liao KP. Enabling phenotypic big data with PheNorm. Journal of the American Medical Informatics Association. 2018 Jan;25(1):54-60. [link]