Predicting breast cancer to improve screening

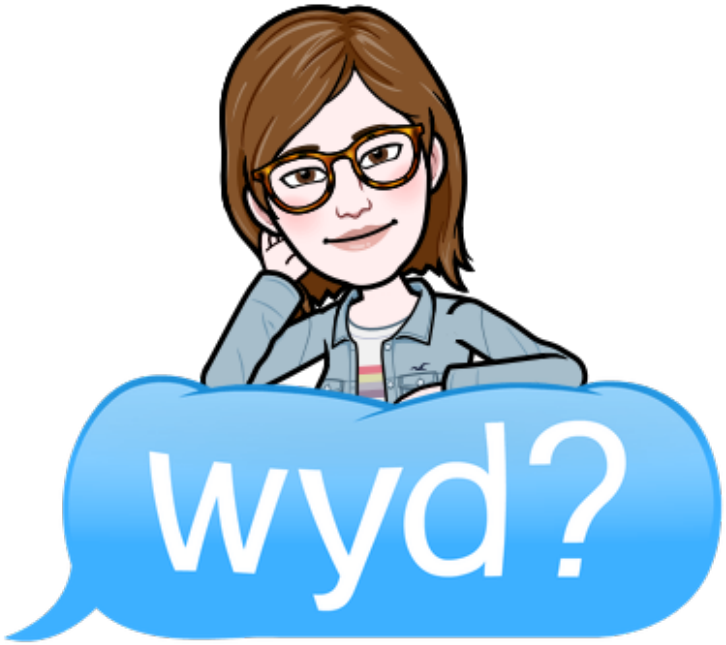
The Women of OHDSI

Introducing the Women of OHDSI (WoO)

Maura Beaton, MS
OHDSI Coordinating Center
Columbia University

Aim of WoO

- Provide a forum where women can share their perspectives, raise concerns and discuss the challenges they face as women working in real-world analytics
- Propose ideas on how the OHDSI community can support women in science and technology
- Support and inspire women to become leaders within the community and their respective fields



Selecting a Study Question

Maura Beaton, MS
OHDSI Coordinating Center
Columbia University

Formulating a Prediction Question

Among patients who **[insert patient cohort]**, which patients will go on to have **[insert outcome of interest]** within **[time window]**?

Example:

Of patients **newly diagnosed with major depressive disorder**, which patients will go on to have **a suicidal event** within **1-year of their diagnosis**?

tduarte Talita

Hi everybody,

My name is Ta
SIDIAP (https

I think it is a g
participate.

I'm not sure I'
are all quite re
health. So I'd
already been

I also have an

Cardiovascula
living in devel
increasing nu
treatments. C
are more likel
postmenopau
women withou
breast cancer
with early rec
Cardiovascula
American Hea

So given that
events among
prevention str

- Among
  diseas

Patrick_Ryan May 19

WoO, I'm glad to see your workgroup coming together to support each other in moving forward a goal to generate reliable evidence.

I support whatever question you ultimately settle on and would be delighted to help in any way I can once you decide on a study.

To add some additional study ideas to the table, here's a type of prediction question that could be informative, for which I think our OHDSI data network could usefully contribute:

The US Preventative Services Task Force recommends regular screening for women for a variety of conditions, including breast, cervical, colorectal (colon) cancers. For each of these screenings, there is some diagnostic procedure performed which can detect the presence of the condition at that time. If a person tests positive, some additional diagnostics and then treatment intervention can be considered; if a person tests negative, the person is recommended to return in some time interval to be retested.

1. Woman aged 30 to 65 are recommended to been screened for cervical cancer every 3-5 years with cervical cytology and.or hrHPV testing. (https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/cervical-cancer-screening2 1)
2. All women aged 50 to 74 are recommended to be screened for breast cancer every 2 years with mammography, but there remains debate about screening mammography when aged 40-49, as it can depend on patient preference toward the benefit-risk tradeoff (https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/breast-cancer-screening1 3)
3. Adults (men or women) aged 50 to 75 are recommended for colorectal cancer screening through multiple methods under different frequency intervals. (https://www.uspreventiveservicestaskforce.org/Page/Document/UpdateSummaryFinal/colorectal-cancer-screening2)

So, while a screening is designed to support immediate detection of disease, my thought it that is also offers a useful moment in time to consider the application of a Patient-Level Prediction model to discuss future risk. In this way, a patient can be educated not just on 'do you have the disease today?' but 'what is the chance you will develop the disease in the next time horizon?'. Having this personalized knowledge may encourage greater adherence to the screening recommendations for followup care.

An example framing of the prediction problem to complement an existing USPSTF screening recommendation would be:

Amongst women aged 40-74 who are undergo a screening mammagraphy who do not have prior breast cancer and are screened negative, which patients will go on to develop breast cancer in the 90d to 3 years following the screening mammagraphy?

May 16

I would like to

ed 65 to 70, which
visit

ireps May 17

Jun 7

owing **child birth**
those who have

ommended for using
ave an elevated risk
ave asked if it's the
a new autoimmune
s that those with
eroid usage), after
utoimmune and an

rd to studies, I see

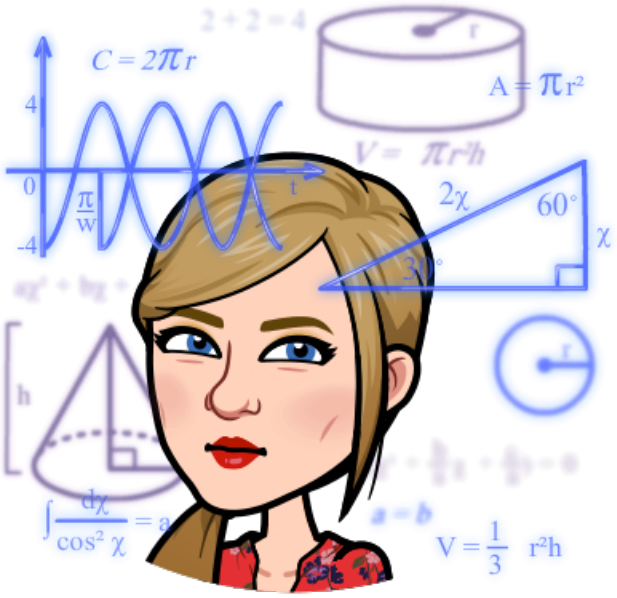
Reply

Question

Amongst **[women aged 40-74 who undergo a screening mammography and who do not have prior breast cancer]**, which patients will go on **[to develop breast cancer]** in the **[90d to 3 years following the screening mammography]**?

Why this question matters

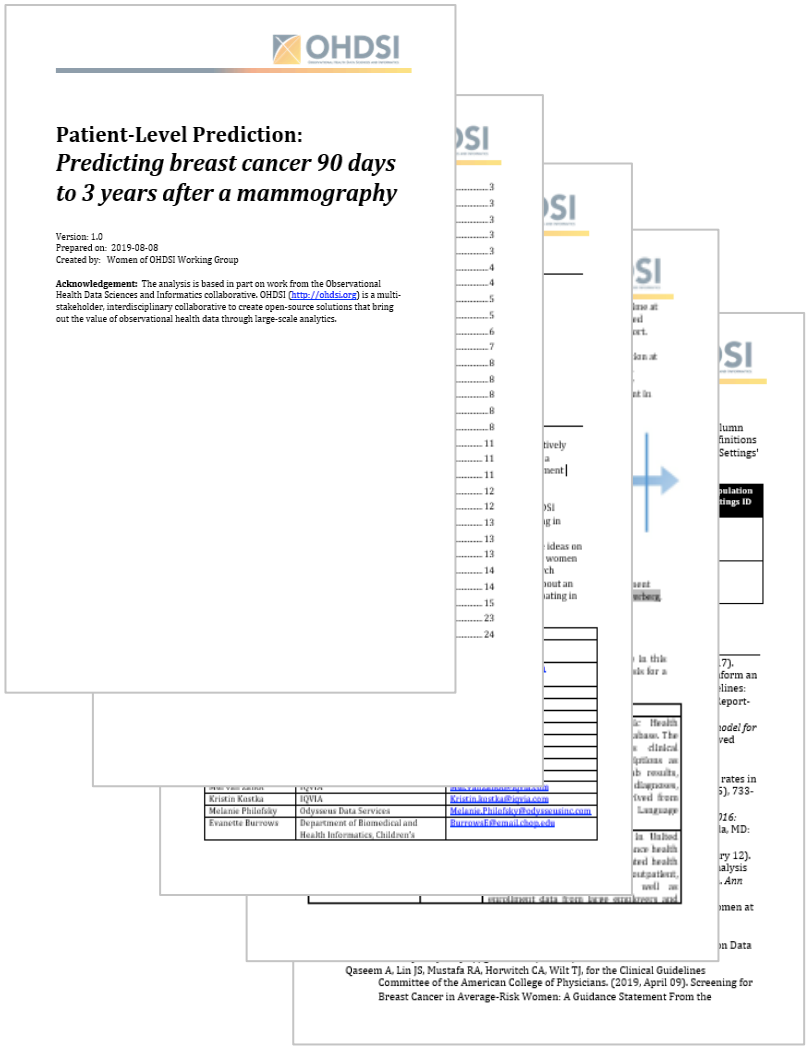
- Demonstrates the risk of developing breast cancer between screenings
  - Encourages patients who underestimate their risk to get regular mammograms
  - Helps patients who overestimate their risk to understand their true likelihood of developing breast cancer
- Ultimately allows patients to make informed, confident decisions about preventative care



Writing a Study Protocol

Kristin Kostka, MPH
Associate Director, OMOP Data Networks
IQVIA

What goes into a Study Protocol



- Responsible Parties
- Objective
- Methods
  - Study Design
  - Data Source(s)
  - Study Populations
  - Statistical Methods
  - Quality Control
- Diagnostics
- Data Analysis Plan
- Strengths & Limitations
- Protection of Human Subjects
- Plans for Disseminating & Communicating Study Results
- References

Study Populations

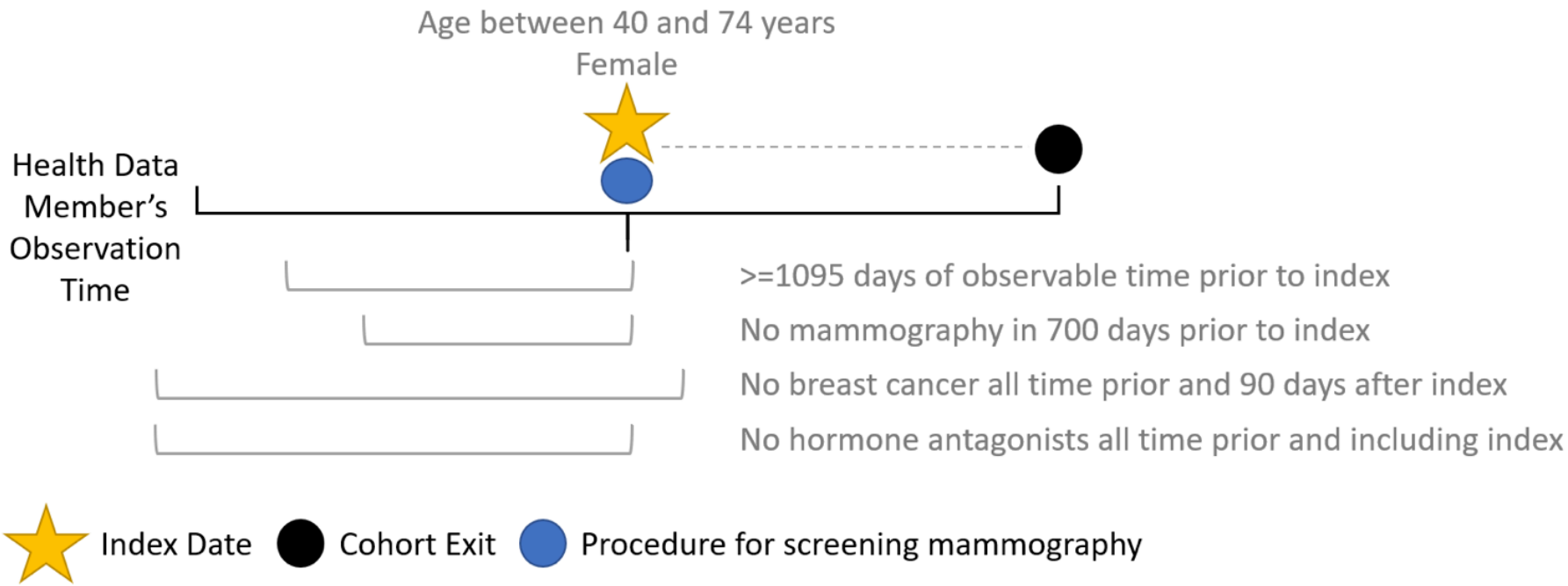
Projects Item	Definition
Target (T)	Women aged 40-74 who are undergo a screening mammography who do not have prior breast cancer.
Outcome (O)	Individuals who develop breast cancer

Time at Risk (TAR) = 90 days after index,

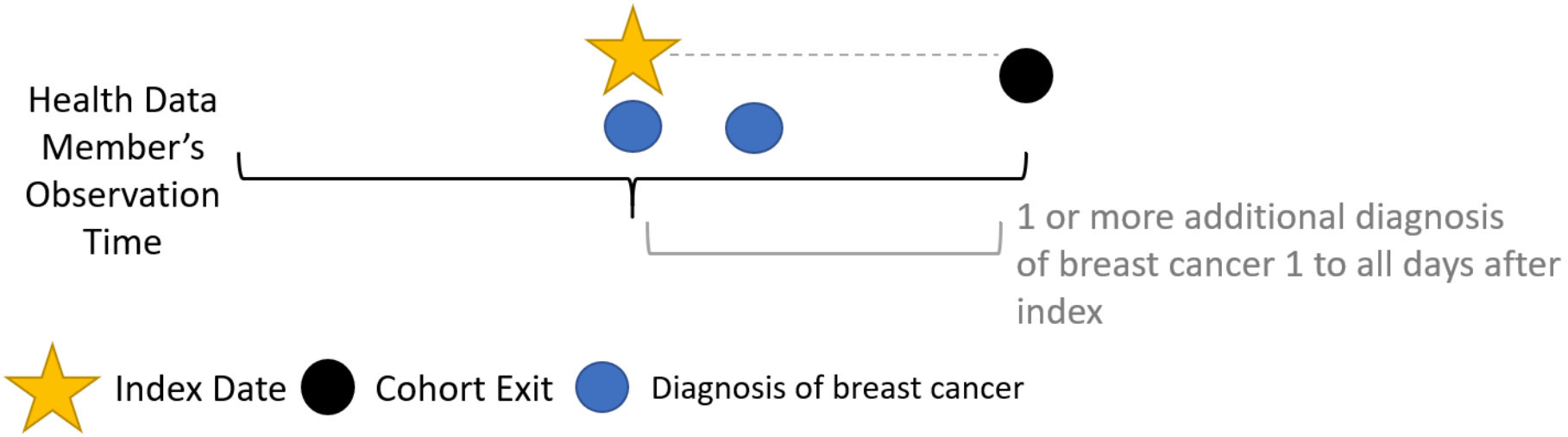
to

1095 days after index

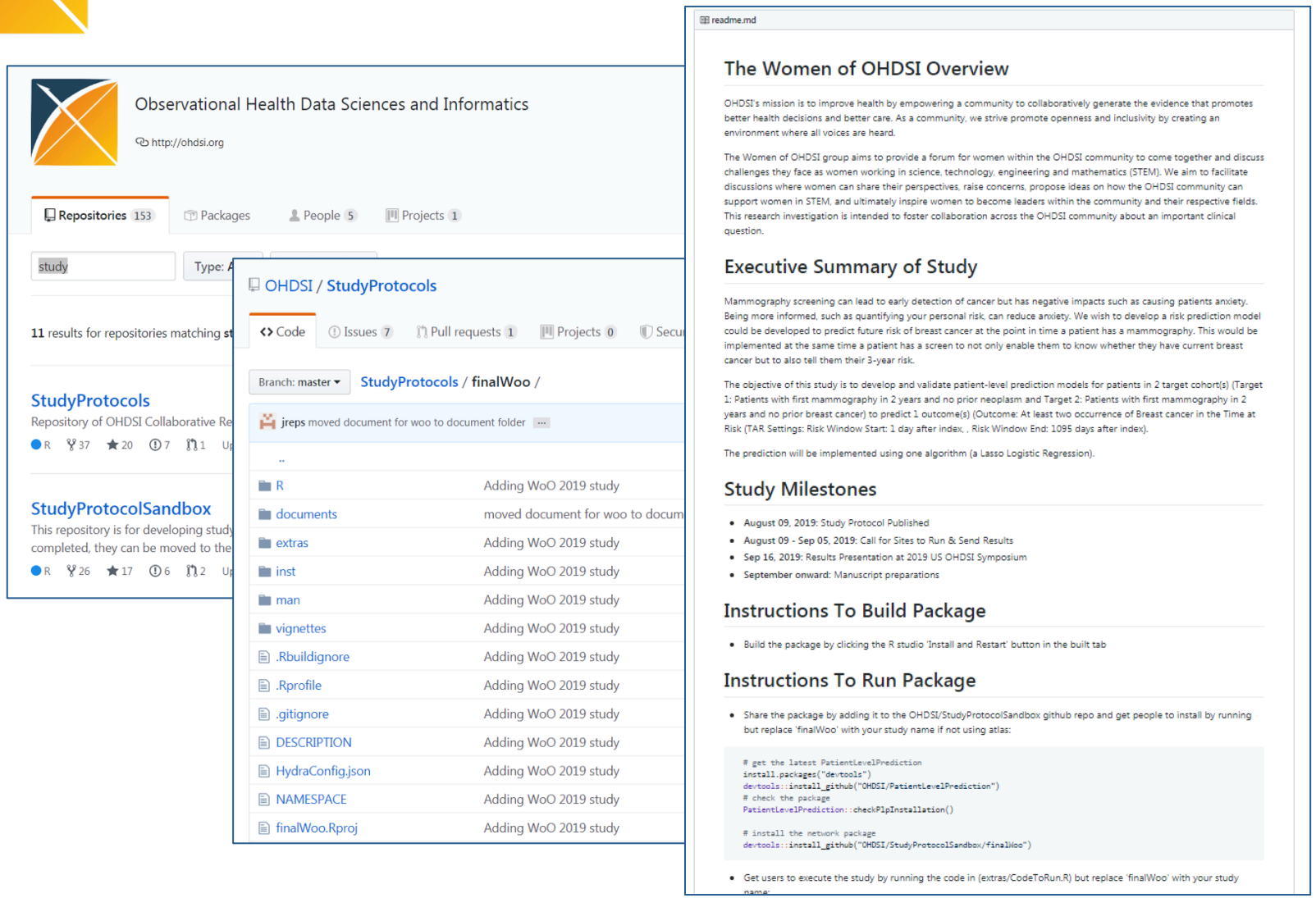
Target Cohort



Outcome Cohort



Publishing to the Community



https://github.com/OHDSI/StudyProtocols/tree/master/finalWoo

Participating Network

Database	Contributor	Description
IBM MarketScan® Commercial Database (CCAE)	Janssen	US commercial claims patients (0-65 years old)
IBM MarketScan® Multi-State Medicaid Database (MDCD)	Janssen	Medicaid enrollees from multiple states
IBM MarketScan® Medicare Supplemental Database (MDCR)	Janssen	Medicare supplemental coverage through privately insured, fee-for-service, point-of-service, or capitated health plans
Optum® De-Identified Clinformatics® Data Mart Database (Optum claims)	Janssen	Primarily representative of US commercial claims patients (0-65 years old) with some Medicare (65+ years old)
Optum® de-identified Electronic Health Record Dataset (Optum EHR)	Janssen	Represents Humedica's EHR medical records database
Columbia University Medical Center Clinical Data Warehouse (CUMC)	Columbia	EHR from the teaching tertiary care hospital
IQVIA LRxDx Open Claims (LRXDX)	IQVIA	Anonymized, pre-adjudicated claims collected from US office-based physicians and specialists
IQVIA Hospital Charge Detail Master (CDM)	IQVIA	Anonymized hospital charge detail masters (CDM) collected from short-term, acute-care and non-federal hospitals
IQVIA US Ambulatory EMR (AmbEMR)	IQVIA	EMR data from US primary care (40%) and speciality practices (60%)
Stanford Medicine Research Data Repository (STaRR)	Stanford	EHR data derived from all patients treated as outpatients and inpatients at Stanford Hospital and Clinics
Regenstrief Institute Indiana Network of Patient Care (INPC)	Regenstrief	Hospitals, physician practices, public health departments, laboratories, radiology and more in the Indiana Network

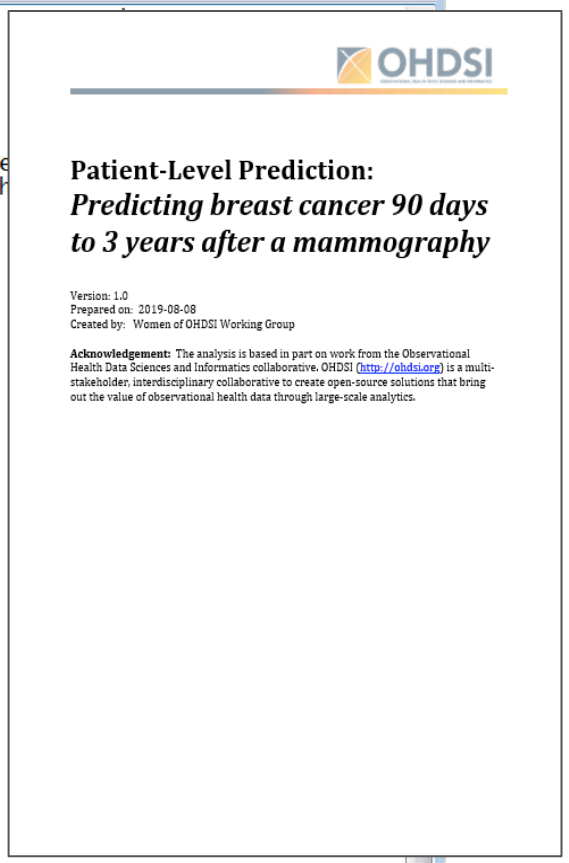
OHDSI Life Hack #31: ATLAS helps write protocols

CodeToRun - Notepad

File Edit Format View Help

```
# Add the database containing the OMOP CDM data
cdmDatabaseSchema <- 'cdm database schema'
# Add a sharebale name for the database containing the OMOP CDM
cdmDatabaseName <- 'a friendly shareable  name for your database
# Add a database with read/write access as this is where the coh
cohortDatabaseSchema <- 'work database schema'

oracleTempSchema <- NULL

# table name where the cohorts will be generated
cohortTable <- 'finalWoOCohort'
#=======================

execute(connectionDetails = connectionDetails,
        cdmDatabaseSchema = cdmDatabaseSchema,
        cdmDatabaseName = cdmDatabaseName,
        cohortDat
        cohortTab
        outputFo
        createPro
        createCohorts = T,
        runAnalyses = T,
        createResultsDoc = T,
        packageResults = T,
        createValidationPackage = T,
        minCellCount= 5,
        createShiny = T,
        createJournalDocument = T,
        analysisIdDocument = 1)

# if you ran execute with: createShiny = T
# Uncomment and run the next line to see the shiny app:
# viewShiny()
```

createProtocol = T,


OHDSI

Patient-Level Prediction:
Predicting breast cancer 90 days to 3 years after a mammography

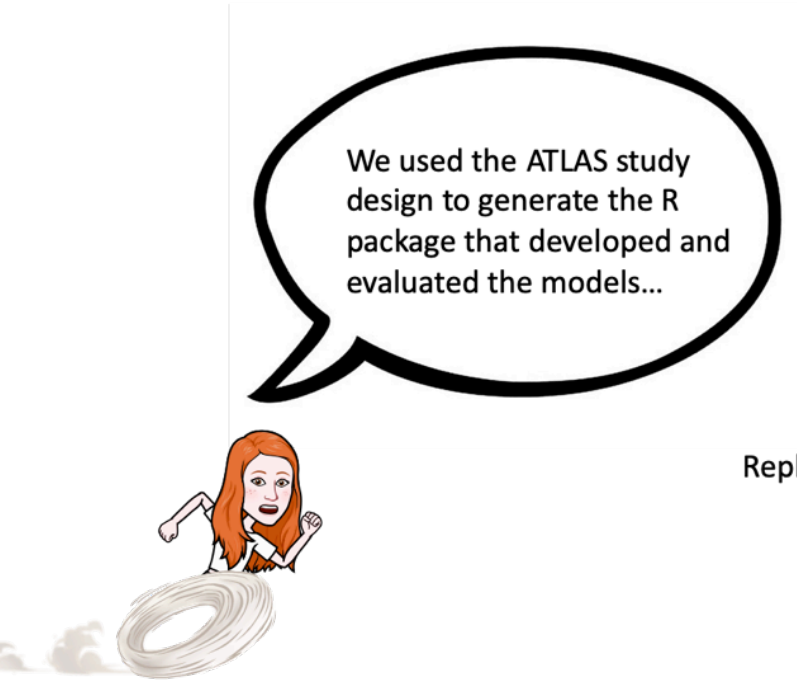
Version: 1.0
Prepared on: 2019-08-08
Created by: Women of OHDSI Working Group

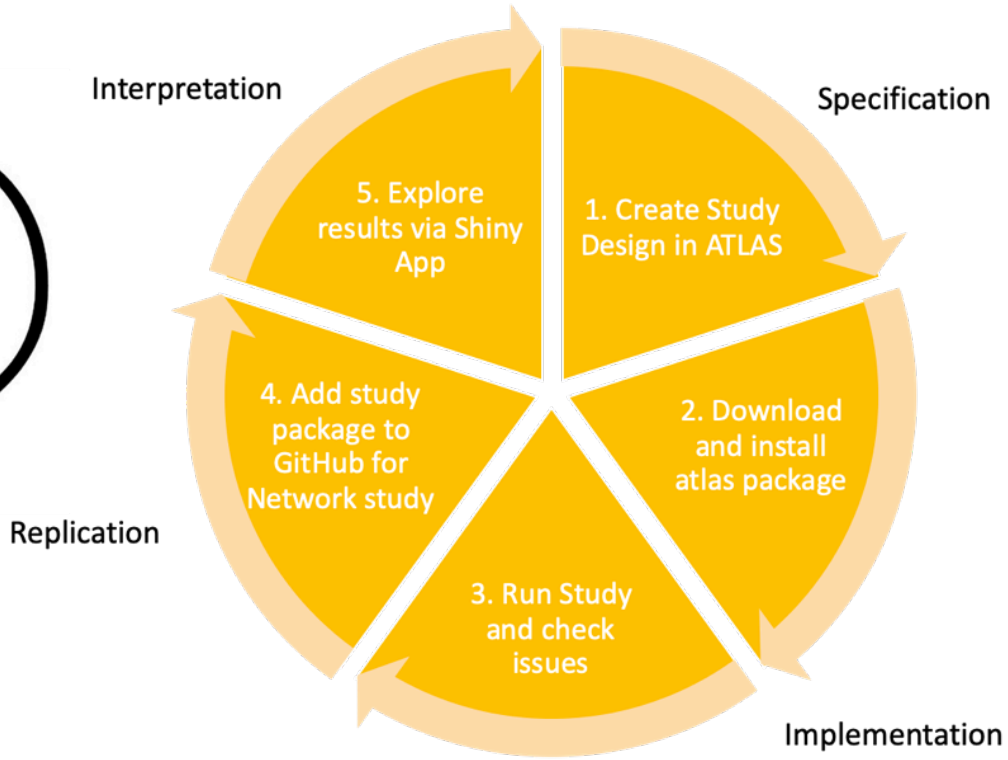
Acknowledgement: The analysis is based in part on work from the Observational Health Data Sciences and Informatics collaborative. OHDSI (http://ohdsi.org) is a multi-stakeholder, interdisciplinary collaborative to create open-source solutions that bring out the value of observational health data through large-scale analytics.

Running an Analysis across the OHDSI Network

Jenna Reps, PhD
Associate Director Epidemiology Analytics
Janssen Research & Development

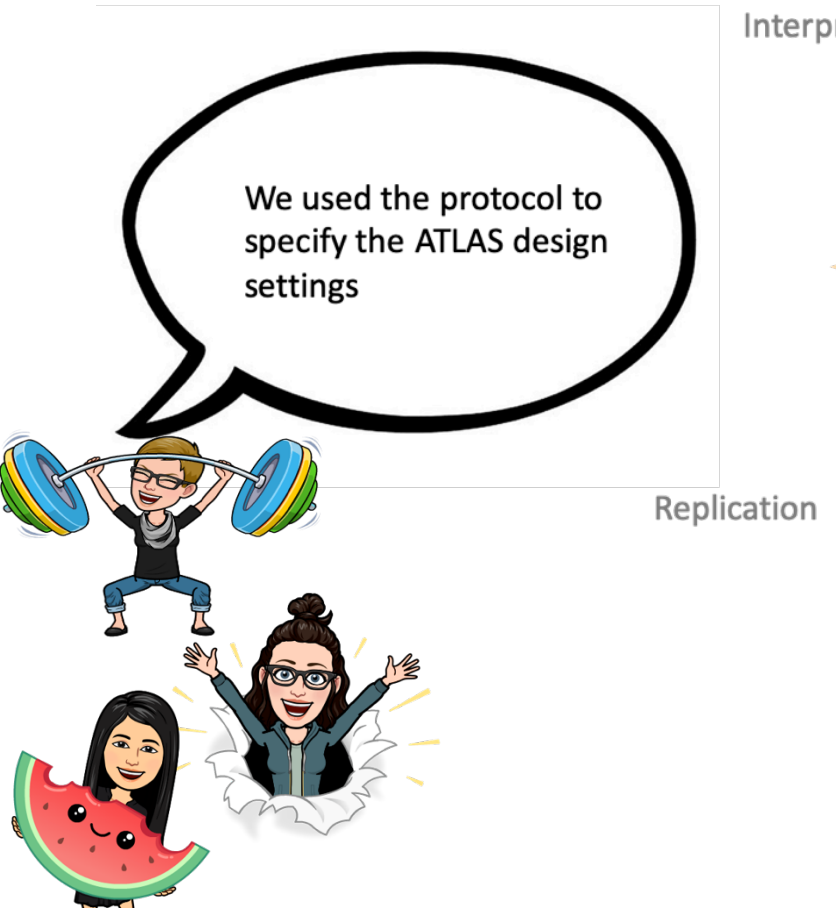
5 Step Process for Prediction Network Study

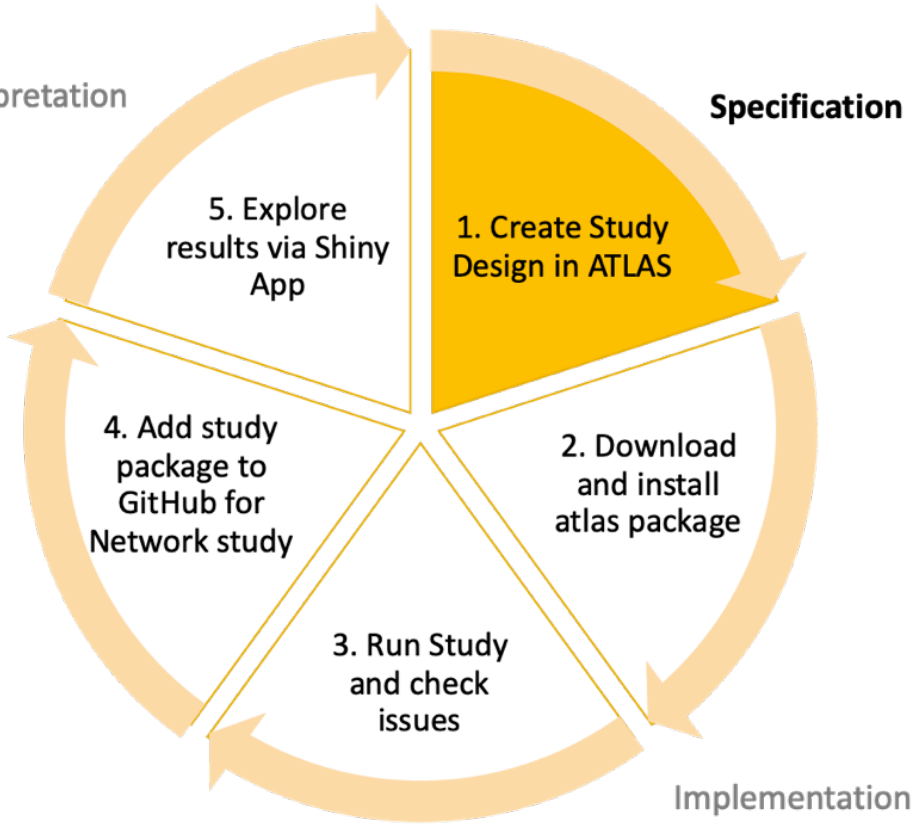




Create your own prediction: http://www.ohdsi.org/web/atlas/#/prediction

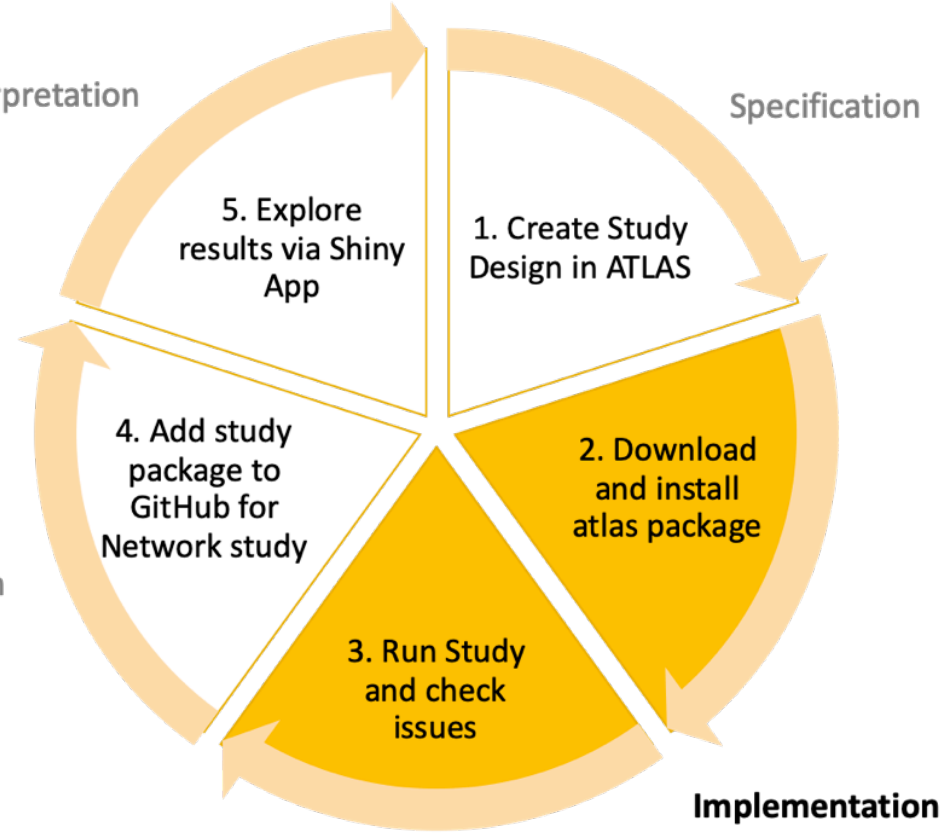
Step 1: Specifying the Prediction





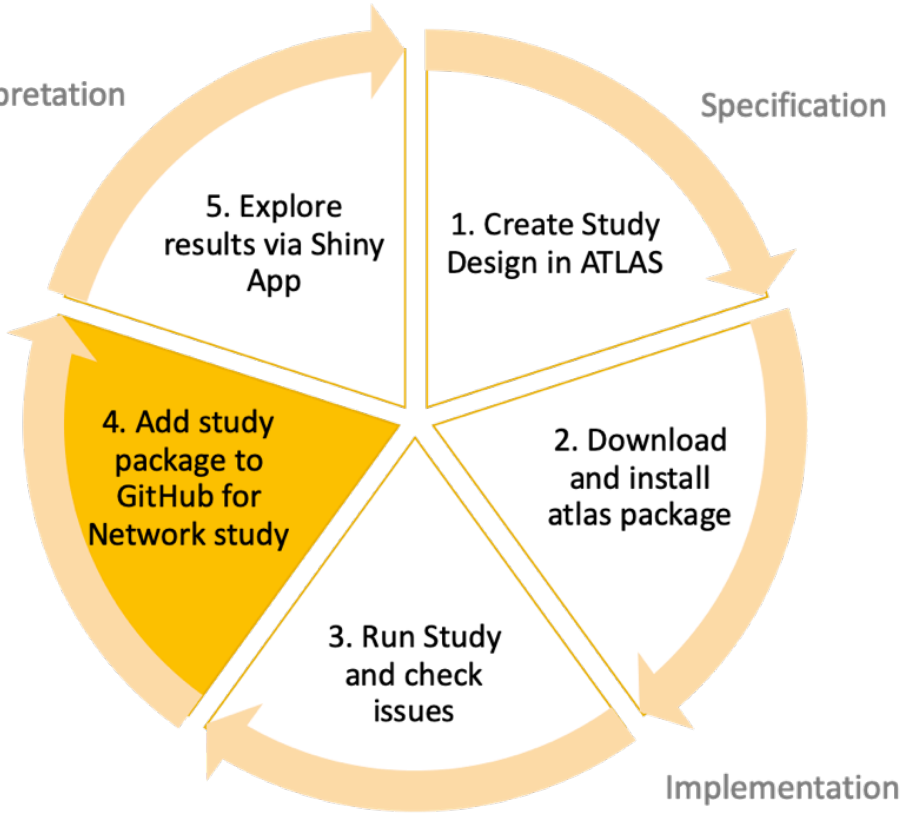
Steps 2 & 3: Initial Development

We set up R
We specified the CDM connection and study databases
We reviewed the model for issues



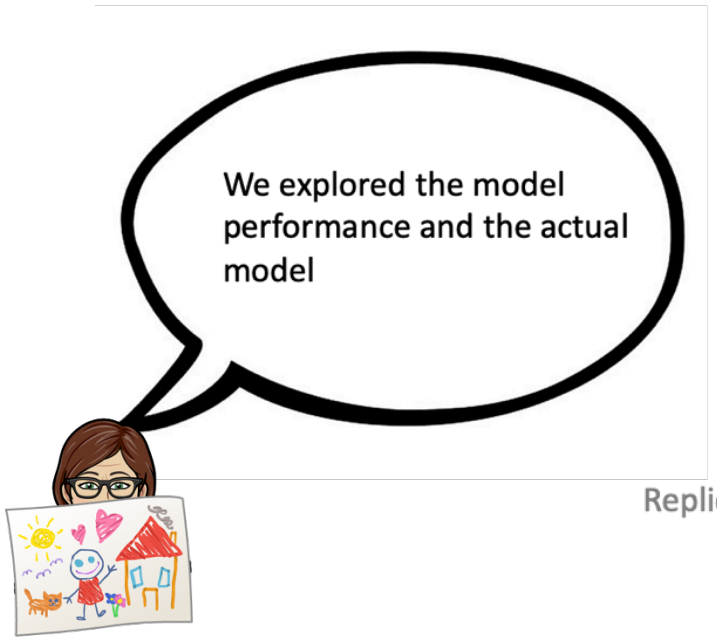
Step 4: Sharing Model

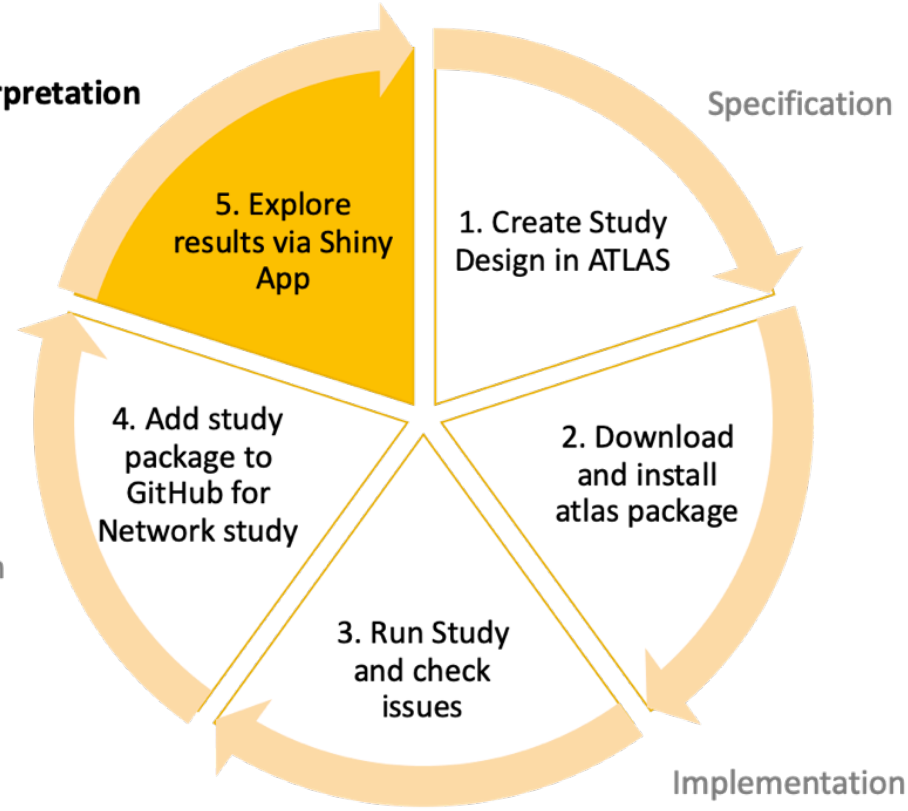
Replication

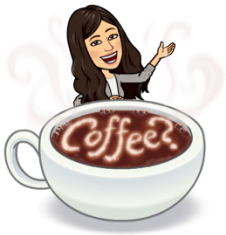


Want to run the study? Go to:
https://github.com/OHDSI/StudyProtocols/tree/master/finalWoo

Step 5: Assessing Model Utility







Want to explore the models?
Go to: http://data.ohdsi.org/WoO2019/

The Results

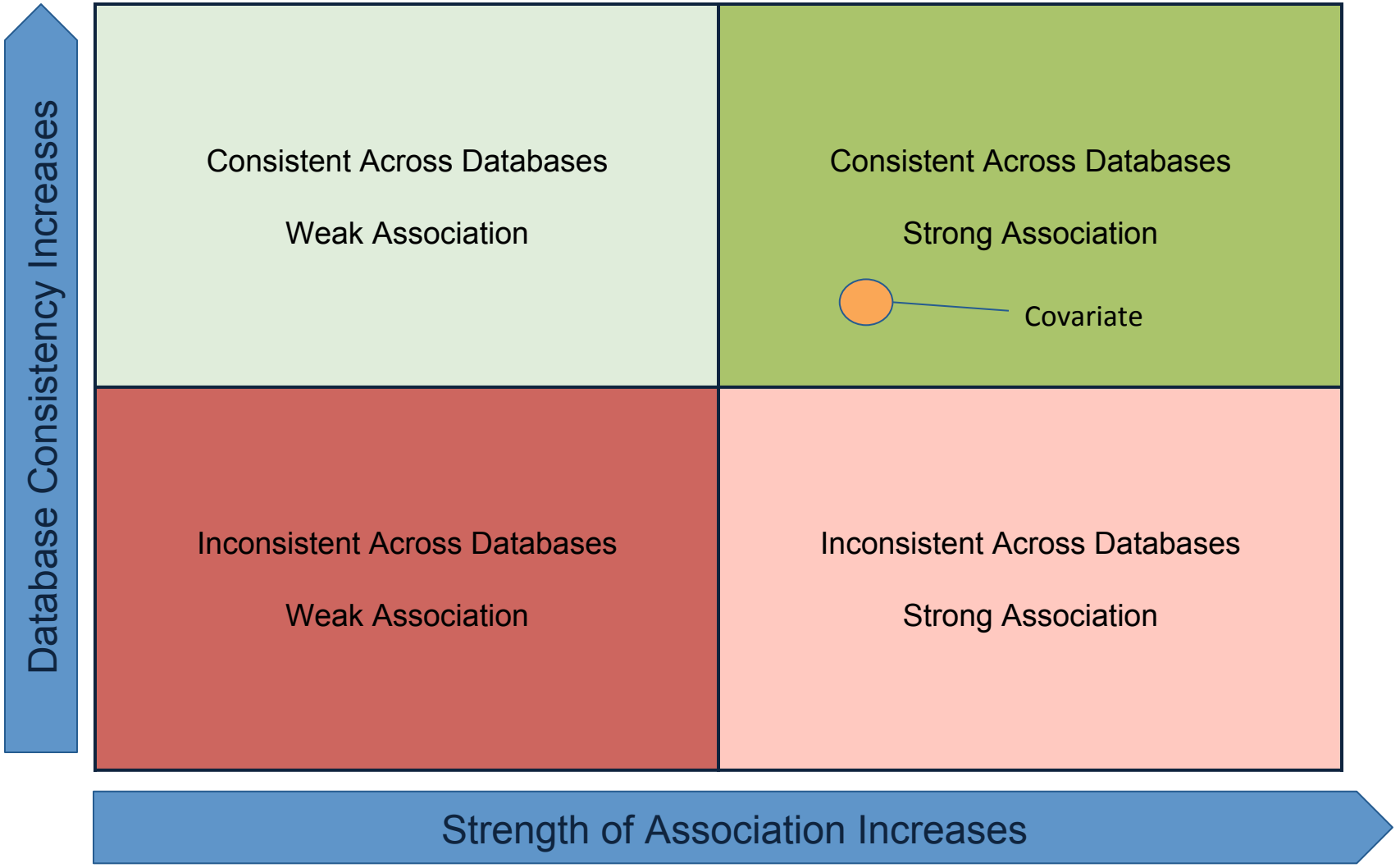
Anna Ostropolets, MD
PhD Student
Columbia University

Area Under the Curve (AUC) & Incidence

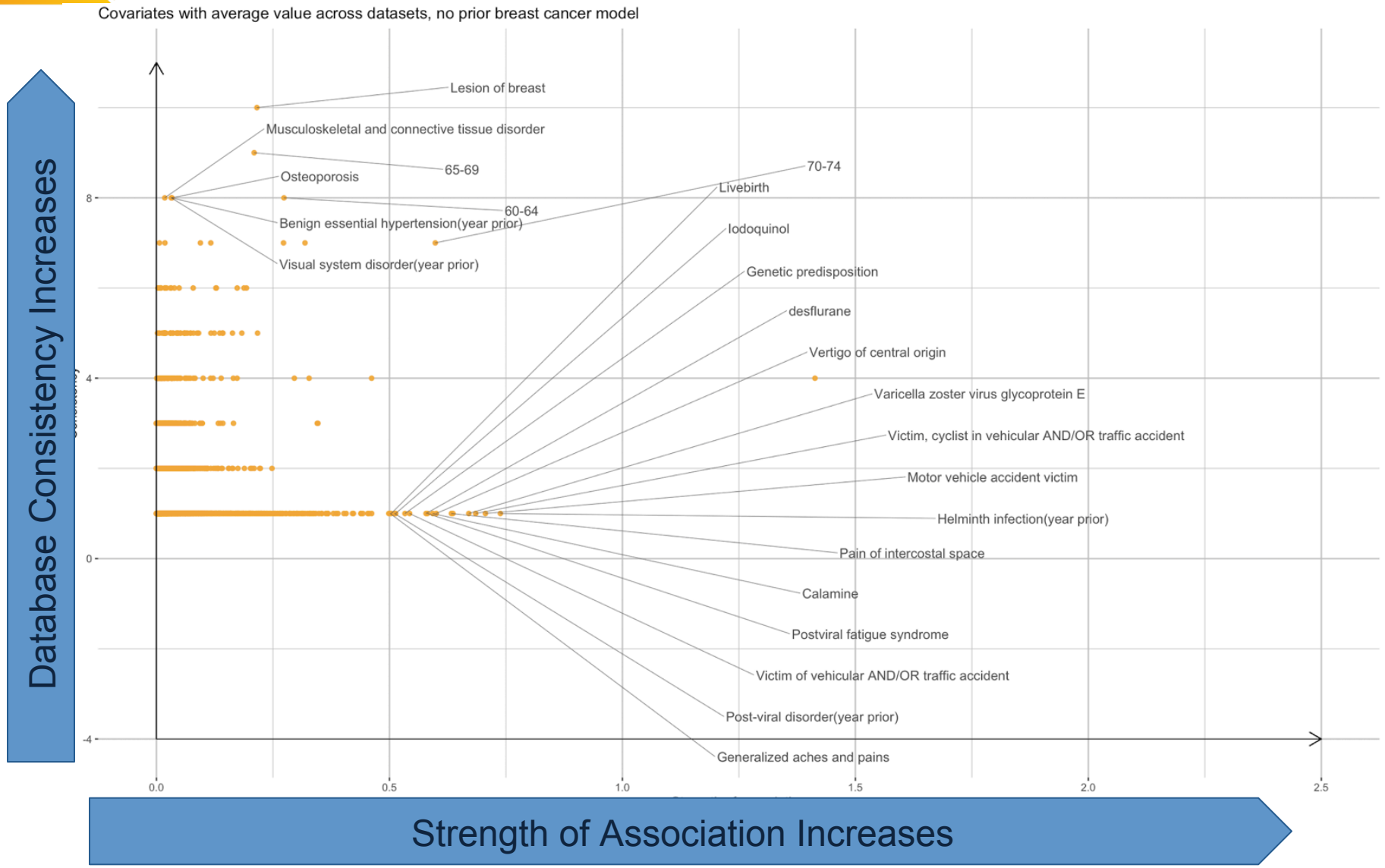
Database	AUC	Incidence Proportion	T size	O size
CCAE	0.62	0.42%	412,572	1,767
MDCD	0.68	0.62%	44,120	274
MDCR	0.57	0.98%	49,782	489
Optum Claims	0.64	0.51%	484,601	2,484
Optum EHR	0.68	1.25%	1,143,599	14,331
CUMC	0.56	0.41%	14,361	59
IQVIA AmbEMR	0.65	0.57%	250,000	1,435
IQVIA LRXDX	0.66	0.57%	250,000	1,425
IQVIA Hospital CDM	0.60	0.35%	250,000	885
	0.61	0.79%	5582	44
f INPC	0.57	0.46%	17958	83

This is important!

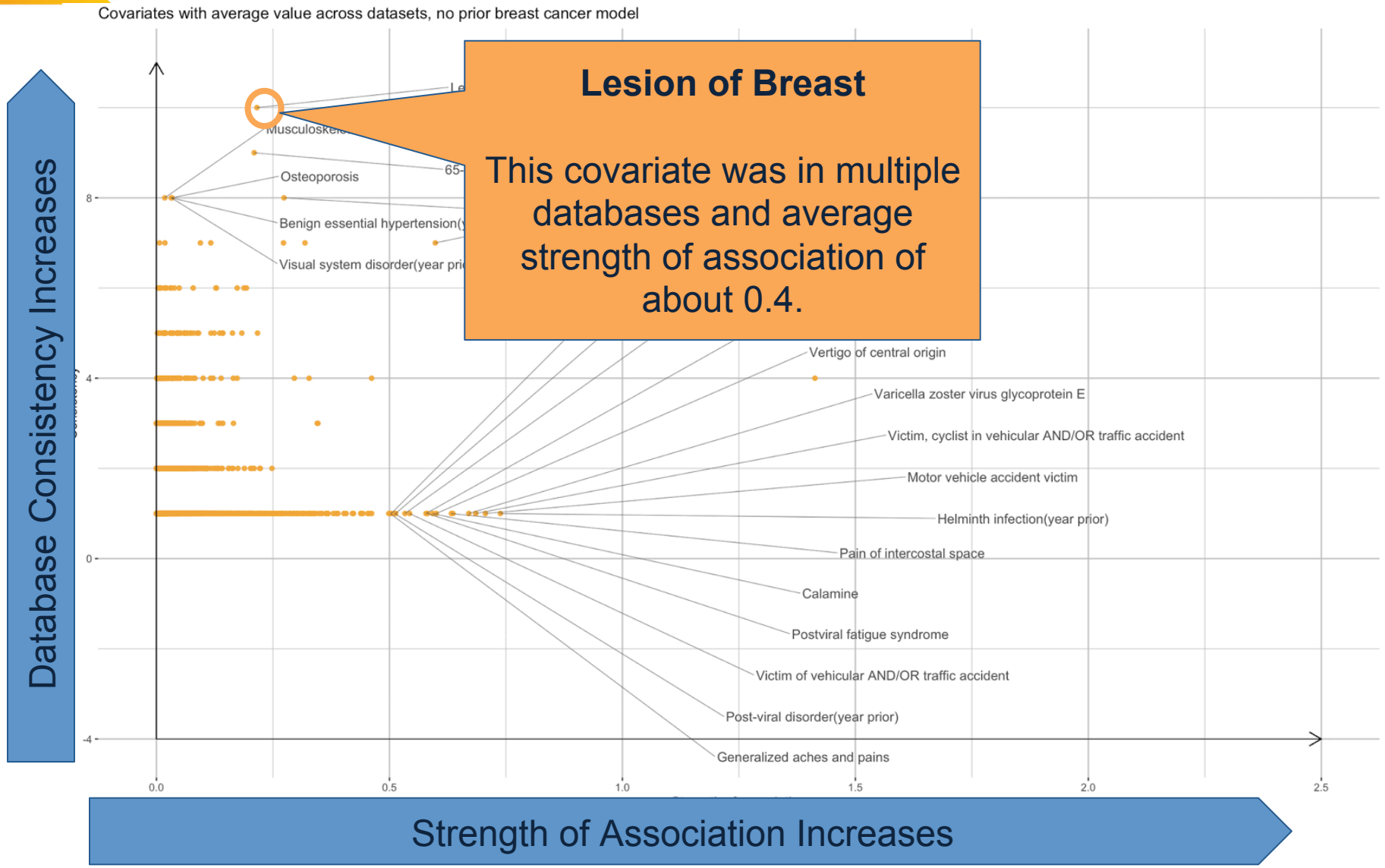
Covariates Across Network



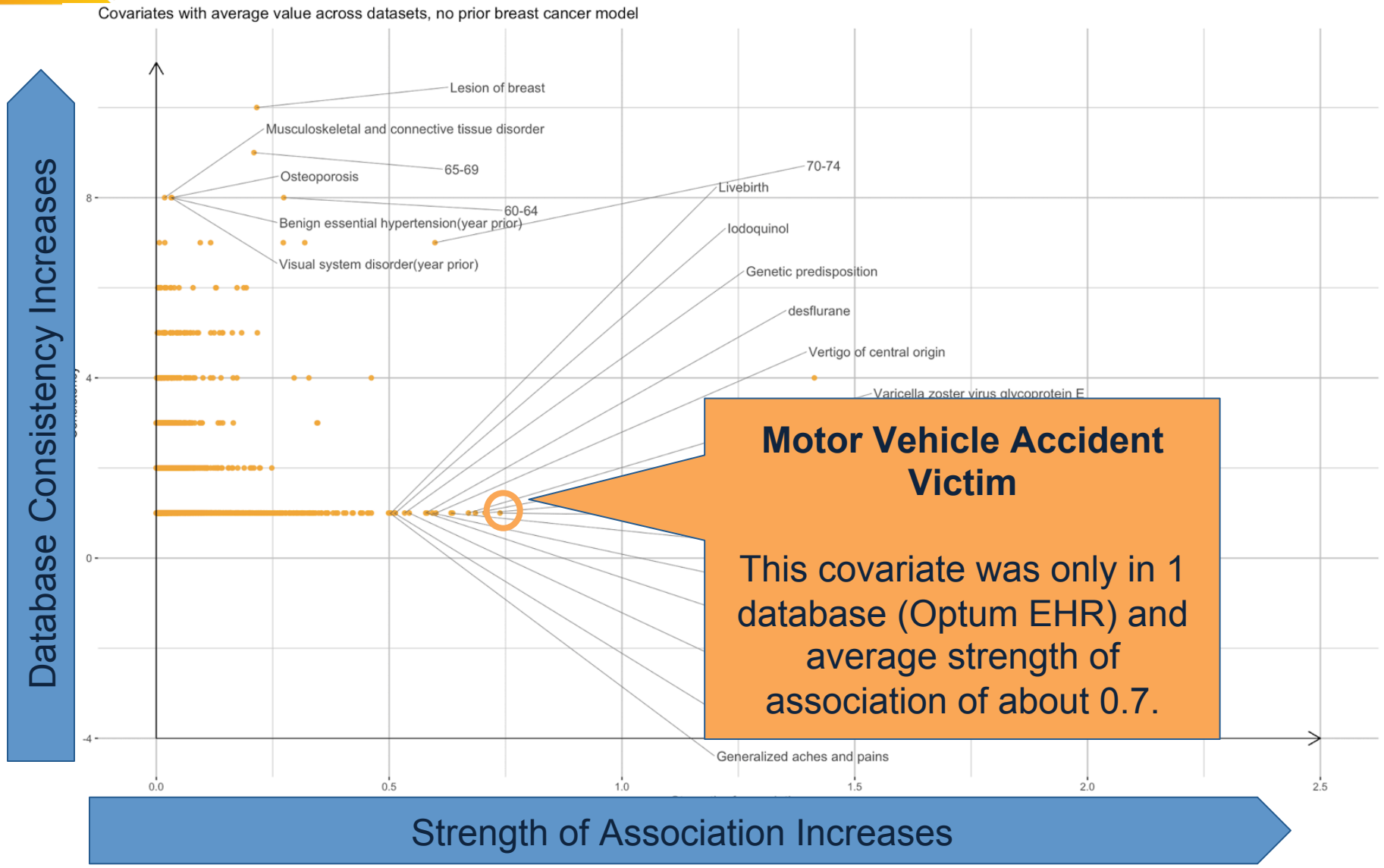
Covariates Across Network



Covariates Across Network



Covariates Across Network



Motor Vehicle Accident Victim

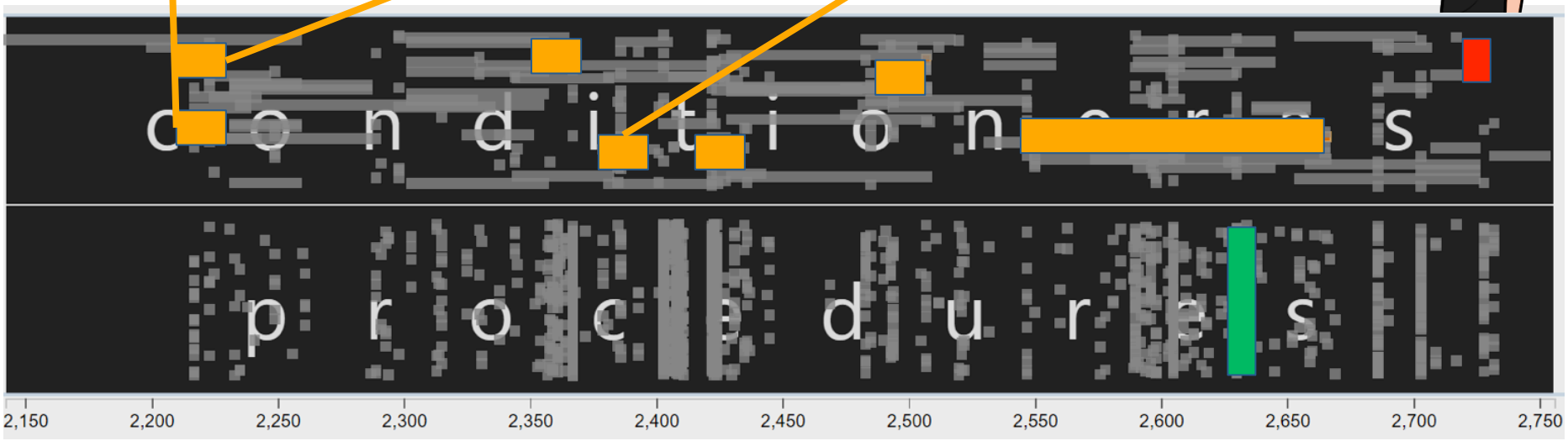
This covariate was only in 1 database (Optum EHR) and average strength of association of about 0.7.

Patient Profiles - True Positive Predicted Risk of 0.96

Helminth infection

Malignant tumor of digestive organ

Generalized aches and pains



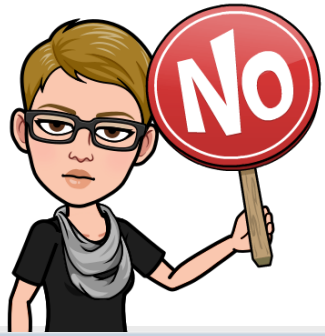
This patient had 25% of the model covariates.

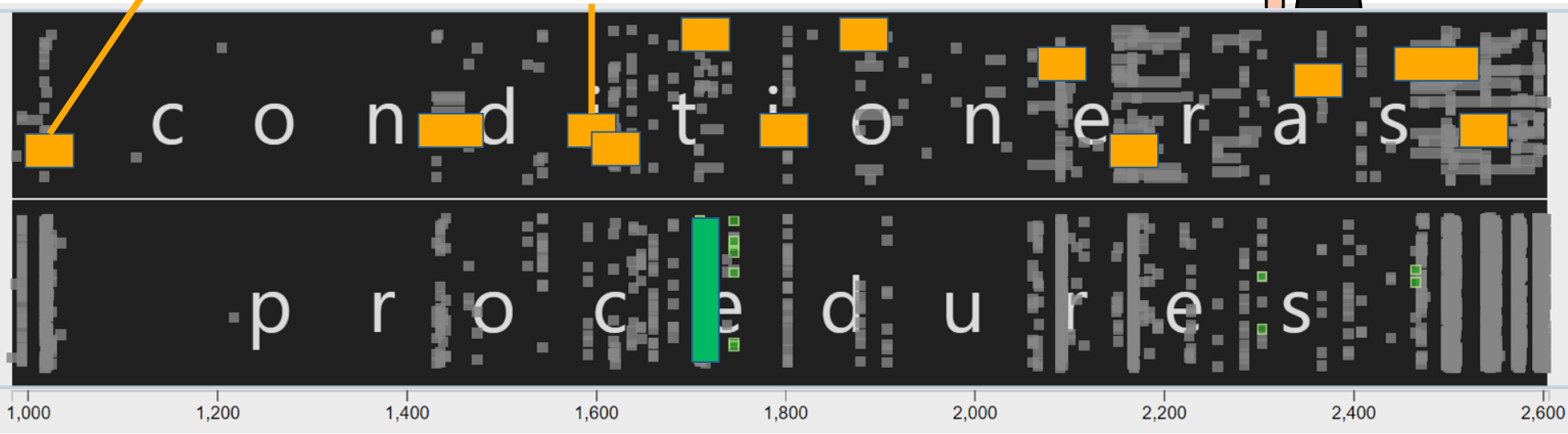
Mammography Risk Factors Breast Cancer

Patient Profiles - False Positive Predicted Risk of 0.90

Neoplasm of digestive system

Generalized aches and pains



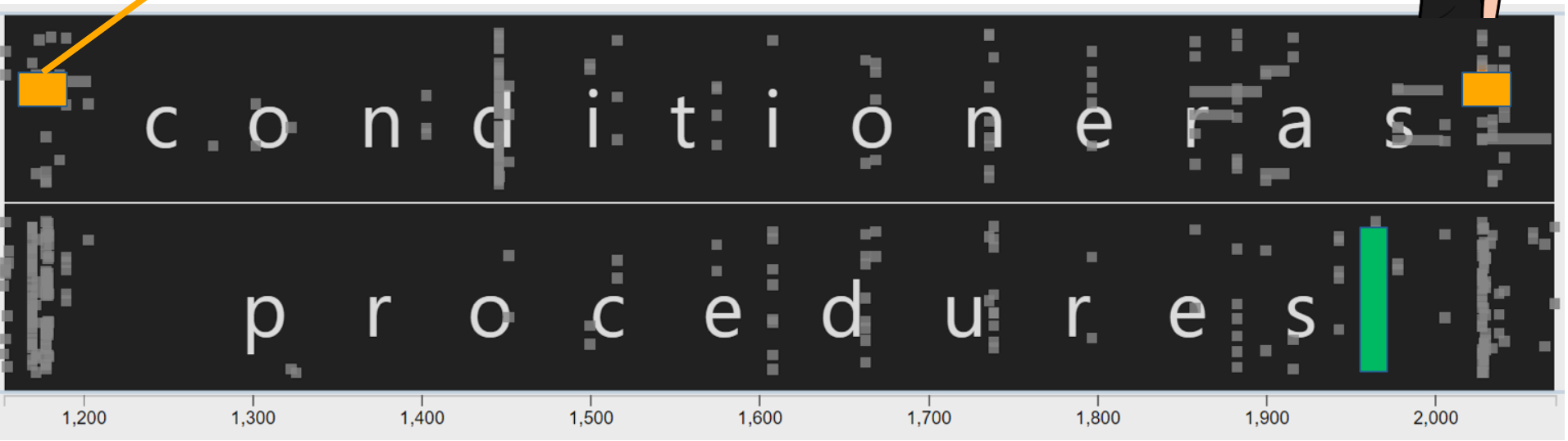


This patient had 16% of the model covariates.
This patient did not develop cancer.

Mammography Risk Factors Breast Cancer

Patient Profiles - True Negative Predicted Risk of 0.00

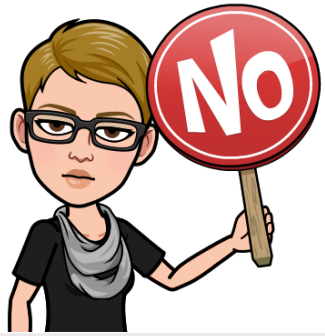
Allergic condition



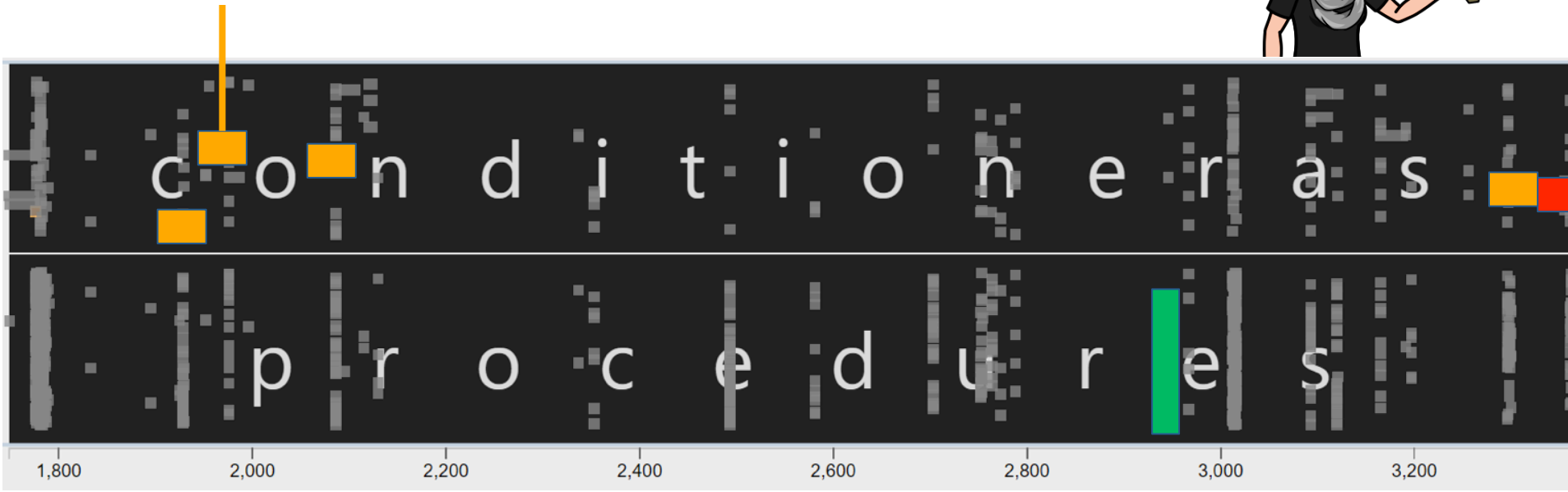
This patient had 9% of the model covariates.
This patient did not develop cancer.

Mammography Risk Factors Breast Cancer

Patient Profiles - False Negative Predicted Risk of 0.00



Allergic condition



This patient had 13% of the model covariates.

Mammography Risk Factors Breast Cancer

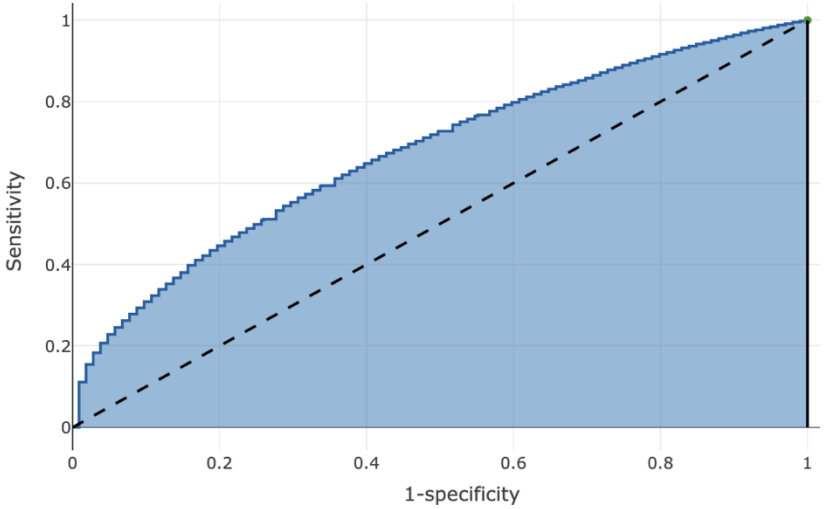
Area Under the Curve (AUC) & Incidence

Database	AUC	Incidence Proportion	T size	O size
CCAE	0.62	0.42%	412,572	1,767
MDCD	0.68	0.62%	44,120	274
MDCR	0.57	0.98%	49,782	489
Optum Claims	0.64	0.51%	484,601	2,484
Optum EHR	0.68	1.25%	1,143,599	14,331
CUMC	0.56	0.41%	14,361	59
IQVIA AmbEMR	0.65			
IQVIA LRXDX	0.66			
IQVIA Hospital CDM	0.60			
STaRR	0.61			
Regenstrief INPC	0.57			

For simplicity, when results are shown we will stick to Optum EHR because it had the best performance and large patient size

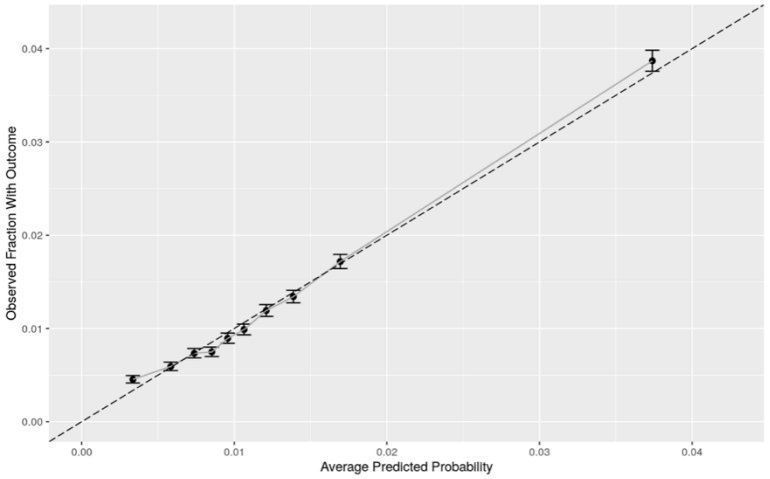
Evaluating the performance of our prediction model

The Optum EHR chose 2,980 covariates in the model

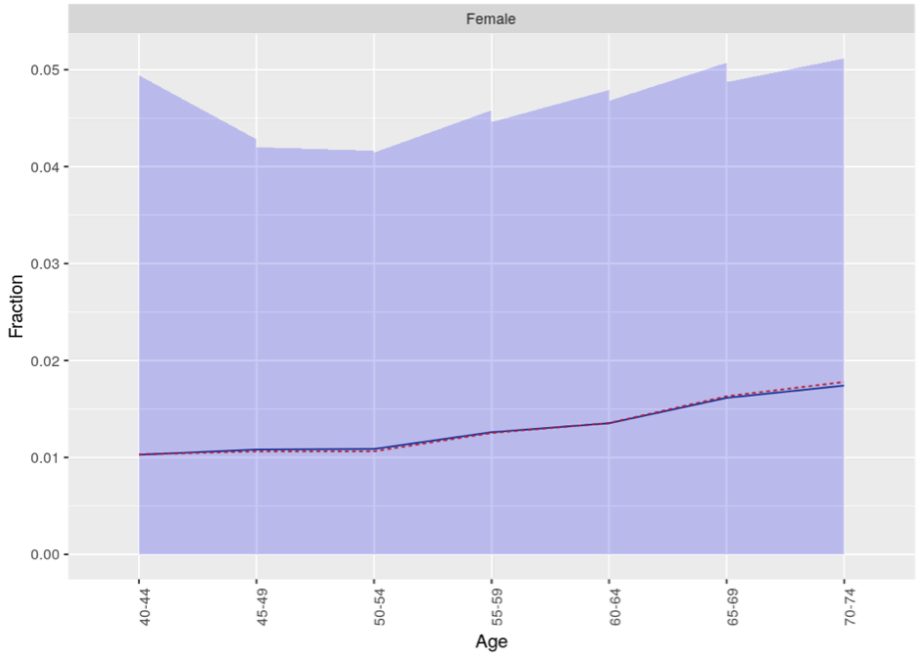


Model Discrimination: Reasonable - although outcome rate of 1 in 100 means high number of false positives

Good Calibration - predicted risk matches observed risk for deciles



Characterizing Risk by Age

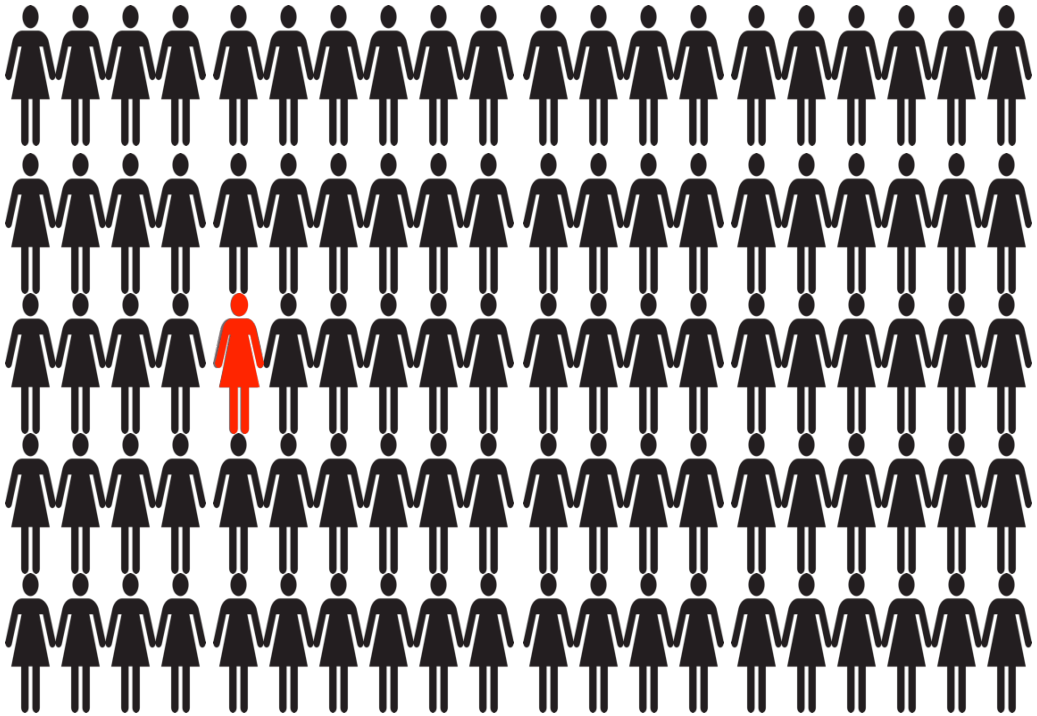


Age Calibration: Good - expected matches observed. Outcome more common in older patients.

Expected (predicted by model)

Observed

Only ~1 in 100 will have breast cancer in next 3 years



No breast cancer in 3 years

Breast cancer within 3 years

Am I the 1?

How to find the 1?

In a world with no patient-level prediction model we currently have three options:

- Do nothing for all (most likely due to rare outcome rate)

	Breast cancer	No breast cancer
Intervene	0	0
Don't intervene	14331	1129268

- Intervene for all

	Breast cancer	No breast cancer
Intervene	14331	1129268
Don't intervene	0	0

- Subjective clinical judgement-based intervention (e.g., 1%)

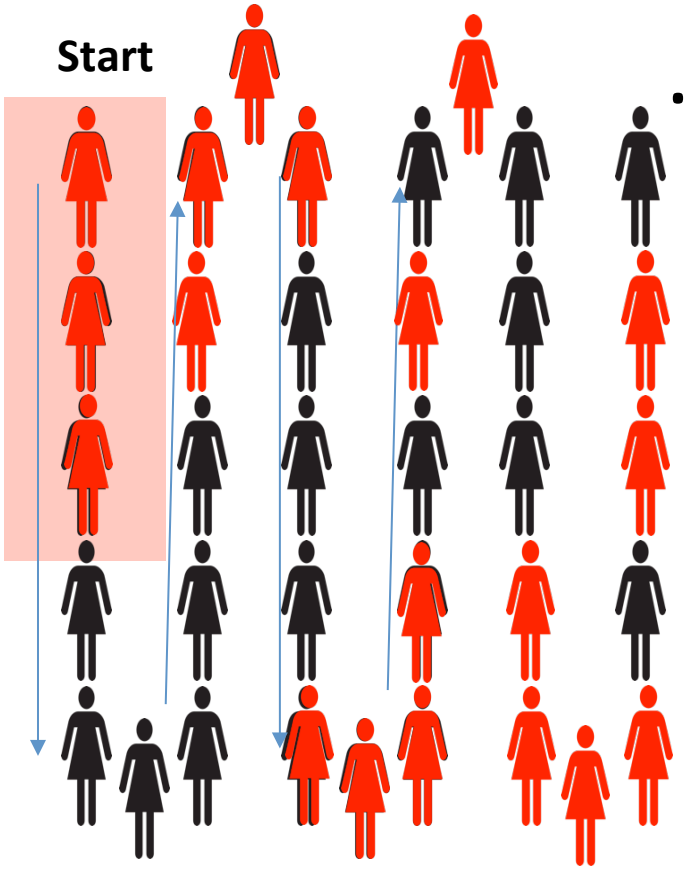
	Breast cancer	No breast cancer
Intervene	143	11293
Don't intervene	14188	1117945

Models got AUC of 0.68 (Optum EHR)

- AUC of 0.68 does not seem great, but…

Models got AUC of 0.68 (Optum EHR)

- AUC of 0.68 does not seem great, but…



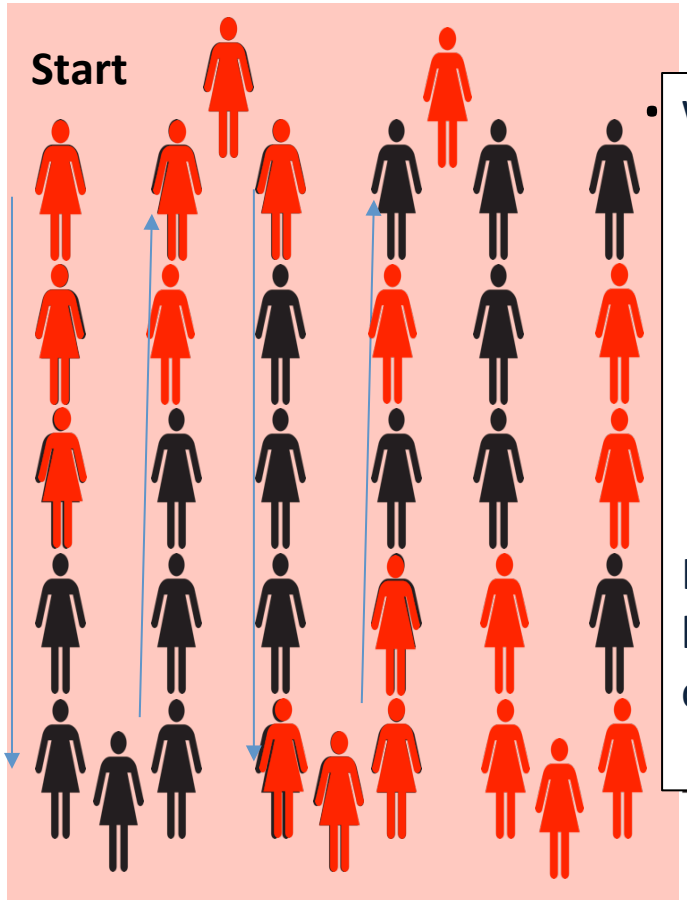
We can identify region of almost certain risk…

	Breast cancer	No breast cancer
Intervene	3	0
Uncertain	14328	1129268

If we only intervene on these patients we can help them without impacting the others at all compared to current care…

Models got AUC of 0.68 (Optum EHR)

- AUC of 0.68 does not seem great, but…



We can identify region of very high risk…

	Breast cancer	No breast cancer
Intervene	19	16
Uncertain	14312	1129252

If we only intervene on these patients we can help them without impacting the others at all compared to current care…

Models got AUC of 0.68 (Optum EHR)

- AUC of 0.68 does not seem great, but…

We can identify region of almost certain no risk…

	Breast cancer	No breast cancer
Uncertain	14331	1128973
Don't intervene	0	295

We could save these patients from the intervention as we know they have a minimal risk and could tell them they have no risk!

Models got AUC of 0.68 (Optum EHR)

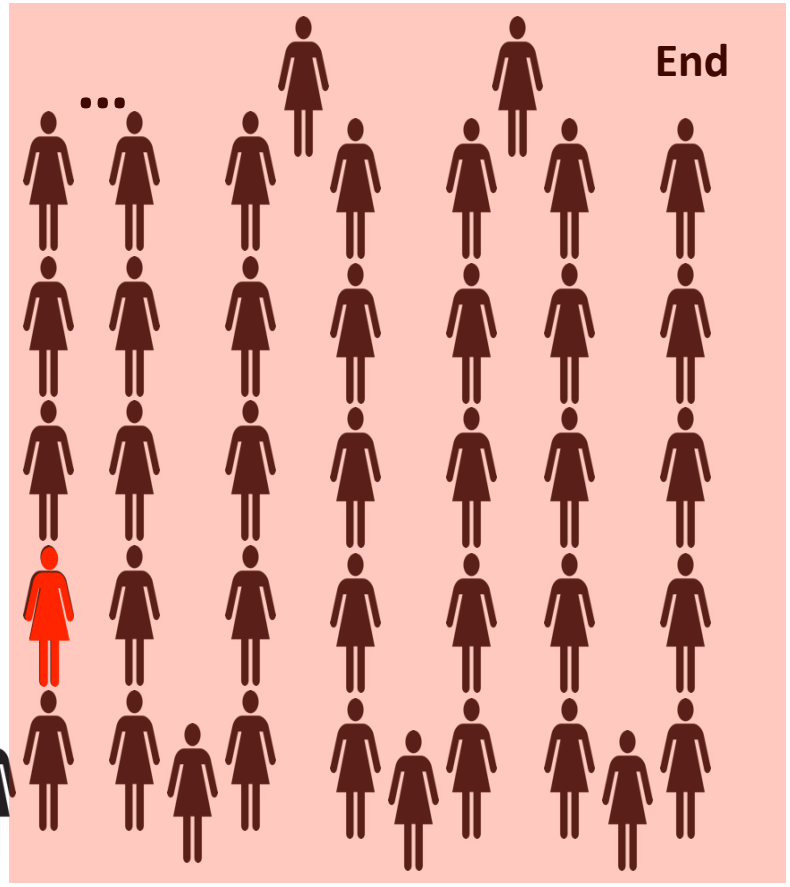
- AUC of 0.68 does not seem great, but…

Start

We can identify region of probable no risk…

	Breast cancer	No breast cancer
Uncertain	14317	1122782
Don't intervene	14	6486

We can save these patients from the intervention as we know they have a minimal risk



Even though we have an AUC of 0.68 (Optum EHR)

We can used the areas the model is certain about:

	Breast cancer	No breast cancer
Intervene	3	0
Don't intervene	0	295
Uncertain	14328	1128973

Almost perfect prediction for 0.03% of patients and the rest get current standard care

We can used the areas the model is confident about:

	Breast cancer	No breast cancer
Intervene	19	16
Don't intervene	14	6486
Uncertain	14298	1122766

Even low discrimination models could have value, even if only helping part of the population...

A model doesn't have to be perfect to be useful

It can find a few with high risk who need closer monitoring

It can find a few with such low risk who are extremely unlikely to develop cancer in 3 years

Could help improve consistency of care

Could help decision making process

We learned a lot...

1. We were able to quantify risk across network
2. We gained insight into variables associated to breast cancer
3. We are able to identify high and low risk subgroups

In future work we will:

- More sensitivity
- Simple Model
- Generate some estimation studies based on our findings
- Write a Paper

Thank you!


Anna
Chun
Clair
Erica
Evanette
Jenna
Jill
Kristin
Laura
Maura
Melanie
Mui
Paola
Rupa
Sarah
Sunny
Zoey
WOMEN OF
8532
OHDSI