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?
Patrick Ryan

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 be 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)

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. 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 is 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 mammography 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 mammography?
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

<table>
<thead>
<tr>
<th>Projects Item</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (T)</td>
<td>Women aged 40-74 who are undergo a screening mammography who do not have prior breast cancer.</td>
</tr>
<tr>
<td>Outcome (O)</td>
<td>Individuals who develop breast cancer</td>
</tr>
</tbody>
</table>

**Time at Risk (TAR) =** 90 days after index, to 1095 days after index
Target Cohort

Age between 40 and 74 years
Female

Health Data
Member’s Observation Time

>=1095 days of observable time prior to index
No mammography in 700 days prior to index
No breast cancer all time prior and 90 days after index
No hormone antagonists all time prior and including index

Index Date  Cohort Exit  Procedure for screening mammography
Outcome Cohort

Health Data
Member’s Observation Time

1 or more additional diagnosis of breast cancer 1 to all days after index

Index Date  Cohort Exit  Diagnosis of breast cancer
Publishing to the Community

https://github.com/OHDSI/StudyProtocols/tree/master/finalWoo
## Participating Network

<table>
<thead>
<tr>
<th>Database</th>
<th>Contributor</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM MarketScan® Commercial Database (CCAE)</td>
<td>Janssen</td>
<td>US commercial claims patients (0-65 years old)</td>
</tr>
<tr>
<td>IBM MarketScan® Multi-State Medicaid Database (MDCD)</td>
<td>Janssen</td>
<td>Medicaid enrollees from multiple states</td>
</tr>
<tr>
<td>IBM MarketScan® Medicare Supplemental Database (MDCR)</td>
<td>Janssen</td>
<td>Medicare supplemental coverage through privately insured,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>fee-for-service, point-of-service, or capitated health plans</td>
</tr>
<tr>
<td>Optum® De-Identified Clininformatics® Data Mart Database</td>
<td>Janssen</td>
<td>Primarily representative of US commercial claims patients (0-65 years old)</td>
</tr>
<tr>
<td>(Optum claims)</td>
<td></td>
<td>with some Medicare (65+ years old)</td>
</tr>
<tr>
<td>Optum® de-identified Electronic Health Record Dataset</td>
<td>Janssen</td>
<td>Represents Humedica’s EHR medical records database</td>
</tr>
<tr>
<td>(Optum EHR)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Columbia University Medical Center Clinical Data Warehouse</td>
<td>Columbia</td>
<td>EHR from the teaching tertiary care hospital</td>
</tr>
<tr>
<td>(CUMC)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>IQVIA LRxDx Open Claims (LRXDX)</td>
<td>IQVIA</td>
<td>Anonymized, pre-adjudicated claims collected from US office-based physicians and specialists</td>
</tr>
<tr>
<td>IQVIA Hospital Charge Detail Master (CDM)</td>
<td>IQVIA</td>
<td>Anonymized hospital charge detail masters (CDM) collected from short-term, acute-care and non-federal hospitals</td>
</tr>
<tr>
<td>IQVIA US Ambulatory EMR (AmbEMR)</td>
<td>IQVIA</td>
<td>EMR data from US primary care (40%) and speciality practices (60%)</td>
</tr>
<tr>
<td>Stanford Medicine Research Data Repository (STaRR)</td>
<td>Stanford</td>
<td>EHR data derived from all patients treated as outpatients and inpatients at Stanford Hospital and Clinics</td>
</tr>
<tr>
<td>Regenstrief Institute Indiana Network of Patient Care (INPC)</td>
<td>Regenstrief</td>
<td>Hospitals, physician practices, public health departments, laboratories, radiology and more in the Indiana Network</td>
</tr>
</tbody>
</table>
OHDSI Life Hack #31: ATLAS helps write protocols

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

We used the ATLAS study design to generate the R package that developed and evaluated the models...
Step 1: Specifying the Prediction

We used the protocol to specify the ATLAS design settings.

1. Create Study Design in ATLAS
2. Download and install atlas package
3. Run Study and check issues
4. Add study package to GitHub for Network study
5. Explore results via Shiny App

Interpretation
Replication
Implementation
Steps 2 & 3: Initial Development

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

1. Create Study Design in ATLAS
2. Download and install atlas package
3. Run Study and check issues
4. Add study package to GitHub for Network study
5. Explore results via Shiny App
Step 4: Sharing Model

Want to run the study? Go to: https://github.com/OHDSI/StudyProtocols/tree/master/finalWoo
Step 5: Assessing Model Utility

We explored the model performance and the actual model.

Want to explore the models?
Go to: [http://data.ohdsi.org/WoO2019/](http://data.ohdsi.org/WoO2019/)
The Results

Anna Ostropolets, MD
PhD Student
Columbia University
## Area Under the Curve (AUC) & Incidence

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

*This is important!*
Covariates Across Network

- Consistent Across Databases
  - Weak Association
  - Strong Association

- Inconsistent Across Databases
  - Weak Association
  - Strong Association

Database Consistency Increases

Strength of Association Increases
Covariates Across Network

- Covariates with average value across datasets, no prior breast cancer model
- Database Consistency Increases
- Strength of Association Increases
Database Consistency Increases

Covariates with average value across datasets, no prior breast cancer model

Lesion of Breast

This covariate was in multiple databases and average strength of association of about 0.4.
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.
Patient Profiles - False Positive
Predicted Risk of 0.90

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

- Neoplasm of digestive system
- Generalized aches and pains

Mammography  Risk Factors  Breast Cancer
Patient Profiles - True Negative Predicted Risk of 0.00

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

- Allergic condition

Mammography  Risk Factors  Breast Cancer
Patient Profiles - False Negative Predicted Risk of 0.00

This patient had 13% of the model covariates.
### Area Under the Curve (AUC) & Incidence

<table>
<thead>
<tr>
<th>Database</th>
<th>AUC</th>
<th>Incidence Proportion</th>
<th>T size</th>
<th>O size</th>
</tr>
</thead>
<tbody>
<tr>
<td>CCAE</td>
<td>0.62</td>
<td>0.42%</td>
<td>412,572</td>
<td>1,767</td>
</tr>
<tr>
<td>MDCD</td>
<td>0.68</td>
<td>0.62%</td>
<td>44,120</td>
<td>274</td>
</tr>
<tr>
<td>MDCR</td>
<td>0.57</td>
<td>0.98%</td>
<td>49,782</td>
<td>489</td>
</tr>
<tr>
<td>Optum Claims</td>
<td>0.64</td>
<td>0.51%</td>
<td>484,601</td>
<td>2,484</td>
</tr>
<tr>
<td>Optum EHR</td>
<td>0.68</td>
<td>1.25%</td>
<td>1,143,599</td>
<td>14,331</td>
</tr>
<tr>
<td>CUMC</td>
<td>0.56</td>
<td>0.41%</td>
<td>14,361</td>
<td>59</td>
</tr>
<tr>
<td>IQVIA AmbEMR</td>
<td>0.65</td>
<td>0.75%</td>
<td>250,000</td>
<td>1,435</td>
</tr>
<tr>
<td>IQVIA LRXDX</td>
<td>0.66</td>
<td>0.57%</td>
<td>250,000</td>
<td>1,425</td>
</tr>
<tr>
<td>IQVIA Hospital CDM</td>
<td>0.60</td>
<td>0.57%</td>
<td>250,000</td>
<td>885</td>
</tr>
<tr>
<td>STaRR</td>
<td>0.61</td>
<td>0.79%</td>
<td>5582</td>
<td>44</td>
</tr>
<tr>
<td>Regenstrief INPC</td>
<td>0.57</td>
<td>0.46%</td>
<td>17,958</td>
<td>83</td>
</tr>
</tbody>
</table>

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)

• Intervene for all

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

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervene</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Uncertain</td>
<td>14328</td>
<td>1129268</td>
</tr>
</tbody>
</table>

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...

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervene</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>Uncertain</td>
<td>14312</td>
<td>1129252</td>
</tr>
</tbody>
</table>

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…

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertain</td>
<td>14331</td>
<td>1128973</td>
</tr>
<tr>
<td>Don’t intervene</td>
<td>0</td>
<td>295</td>
</tr>
</tbody>
</table>

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…

We can identify region of probable no risk…

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uncertain</td>
<td>14317</td>
<td>1122782</td>
</tr>
<tr>
<td>Don’t intervene</td>
<td>14</td>
<td>6486</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervene</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Don’t intervene</td>
<td>0</td>
<td>295</td>
</tr>
<tr>
<td>Uncertain</td>
<td>14328</td>
<td>1128973</td>
</tr>
</tbody>
</table>

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:

<table>
<thead>
<tr>
<th></th>
<th>Breast cancer</th>
<th>No breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intervene</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>Don’t intervene</td>
<td>14</td>
<td>6486</td>
</tr>
<tr>
<td>Uncertain</td>
<td>14298</td>
<td>1122766</td>
</tr>
</tbody>
</table>

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!