**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.

**Table of Contents**

[1. List of Abbreviations 3](#_Toc16170869)

[2. Responsible Parties 3](#_Toc16170870)

[3. Executive Summary 3](#_Toc16170871)

[4. Rational & Background 3](#_Toc16170872)

[5. Objective 3](#_Toc16170873)

[6. Methods 4](#_Toc16170874)

[6.1. Study Design 4](#_Toc16170875)

[6.2. Data Source(s) 5](#_Toc16170876)

[6.3. Study Populations 5](#_Toc16170877)

[6.4. Statistical Analysis Method(s) 6](#_Toc16170878)

[6.5. Quality Control 7](#_Toc16170879)

[6.6. Tools 8](#_Toc16170880)

[7. Diagnostics 8](#_Toc16170881)

[8. Data Analysis Plan 8](#_Toc16170882)

[8.1. Algorithm Settings 8](#_Toc16170883)

[8.2. Covariate Settings 8](#_Toc16170884)

[8.3. Model Development & Evaluation 11](#_Toc16170885)

[8.4. Analysis Execution Settings 11](#_Toc16170886)

[9. Strengths & Limitations 11](#_Toc16170887)

[10. Protection of Human Subjects 12](#_Toc16170888)

[11. Plans for Disseminating & Communicating Study Results 12](#_Toc16170889)

[12. Tables & Figures 13](#_Toc16170890)

[12.1. Incidence Rate of Target & Outcome 13](#_Toc16170891)

[12.2. Characterization 13](#_Toc16170892)

[13. Appendices 14](#_Toc16170893)

[13.1. Study Generation Version Information 14](#_Toc16170894)

[13.2. Code List 15](#_Toc16170895)

[13.3. Complete Analysis List 23](#_Toc16170896)

[14. References 24](#_Toc16170897)

# List of Abbreviations

| Abbreviation | Phrase |
| --- | --- |
| AUC | Area Under the Receiver Operating Characteristic Curve |
| CDM | Common Data Model |
| O | Outcome Cohort |
| OHDSI | Observational Health Data Sciences & Informatics |
| OMOP | Observational Medical Outcomes Partnership |
| T | Target Cohort |
| TAR | Time at Risk |

# Responsible Parties

OHDSI's mission is to improve health by empowering a community to collaboratively generate the evidence that promotes better health decisions and better care. As a community, we strive promote openness and inclusivity by creating an environment where all voices are heard.

The Women of OHDSI group aims to provide a forum for women within the OHDSI community to come together and discuss challenges they face as women working in science, technology, engineering and mathematics (STEM). We aim to facilitate discussions where women can share their perspectives, raise concerns, propose ideas on how the OHDSI community can support women in STEM, and ultimately inspire women to become leaders within the community and their respective fields. This research investigation is intended to foster collaboration across the OHDSI community about an important clinical question. Below are the affiliated individuals who are participating in this Women of OHDSI study:

|  |  |  |
| --- | --- | --- |
| **Author** | **Affiliation** | **Email** |
| Maura Beaton | Department of Biomedical Informatics, Columbia University | [beaton@ohdsi.org](mailto:beaton@ohdsi.org) |
| Anna Ostropolets | Department of Biomedical Informatics, Columbia University | [aostropolets@gmail.com](mailto:aostropolets@gmail.com) |
| Erica A. Voss | Janssen Research & Development | [evoss3@its.jnj.com](mailto:evoss3@its.jnj.com) |
| Jenna Reps | Janssen Research & Development | [jreps@its.jnj.com](mailto:jreps@its.jnj.com) |
| Jill Hardin | Janssen Research & Development | [jhardi10@its.jnj.com](mailto:jhardi10@its.jnj.com) |
| Laura Hester | Janssen Research & Development | [lhester@its.jnj.com](mailto:lhester@its.jnj.com) |
| Rupa Makadia | Janssen Research & Development | [rmakadia@its.jnj.com](mailto:rmakadia@its.jnj.com) |
| Clair Blacketer | Janssen Research & Development | [mblacke@its.jnj.com](mailto:mblacke@its.jnj.com) |
| Paola Saroufim | Case Western Reserve University | [pxs503@case.edu](mailto:pxs503@case.edu) |
| Sarah Seager | IQVIA | [sarah.seager@iqvia.com](mailto:sarah.seager@iqvia.com) |
| Mui van Zandt | IQVIA | [Mui.VanZandt@iqvia.com](mailto:Mui.VanZandt@iqvia.com) |
| Kristin Kostka | IQVIA | [Kristin.kostka@iqvia.com](mailto:Kristin.kostka@iqvia.com) |
| Melanie Philofsky | Odysseus Data Services | [Melanie.Philofsky@odysseusinc.com](mailto:Melanie.Philofsky@odysseusinc.com) |
| Evanette Burrows | Department of Biomedical and Health Informatics, Children’s Hospital of Philadelphia | [BurrowsE@email.chop.edu](mailto:BurrowsE@email.chop.edu) |
| Xinzhuo (Zoey) Jiang | Department of Biomedical Informatics, Columbia University | [xj2193@cumc.columbia.edu](mailto:xj2193@cumc.columbia.edu) |
| Noemie Elhadad | Department of Biomedical Informatics, Columbia University | [noemie@gmail.com](mailto:noemie@gmail.com) |
| Chunhua Weng | Department of Biomedical Informatics, Columbia University | [cw2384@cumc.columbia.edu](mailto:cw2384@cumc.columbia.edu) |
| Chun Yee Lau | Department of Biomedical Informatics, Columbia University | [cl3777@cumc.columbia.edu](mailto:cl3777@cumc.columbia.edu) |

**Study Timeline:**

August 09, 2019: Study Protocol Published

August 09 - Sep 05, 2019: Call for Sites to Run & Send Results

Sep 16, 2019: Results Presentation at 2019 US OHDSI Symposium

Late September onward: Manuscript preparations

# Executive Summary

Mammography screening can lead to early detection of cancer but has negative impacts such as causing patients anxiety. Being more informed, such as quantifying your personal risk, can reduce anxiety. We wish to develop a risk prediction model could be developed to predict future risk of breast cancer at the point in time a patient has a mammography. This would be implemented at the same time a patient has a screen to not only enable them to know whether they have current breast cancer but to also tell them their 3-year risk.

The objective of this study is to develop and validate patient-level prediction models for patients in 2 target cohort(s) (Target 1: Patients with first mammography in 2 years and no prior neoplasm and Target 2: Patients with first mammography in 2 years and no prior breast cancer) to predict 1 outcome(s) (Outcome: At least two occurrence of Breast cancer in the Time at Risk (TAR Settings: Risk Window Start: 1 day after index, , Risk Window End: 1095 days after index).

The prediction will be implemented using one algorithm (a Lasso Logistic Regression).

# Rationale & Background

Breast cancer is the fourth leading cause of cancer death in the United States, with 1 in 8 US women developing invasive breast cancer throughout her lifetime (Howlader, Noone, & al, 2019). The American College of Physicians (ACP) recommends women between the ages of 50-74 consider a bi-annual mammography as it reduces breast cancer-associated mortality (Qaseem A, Lin JS, Mustafa RA, Horwitch CA, Wilt TJ, for the Clinical Guidelines Committee of the American College of Physicians, 2019). There is no consensus among the current guidelines on how to approach average-risk patients aged 40 to 49 years. Most of them suggest that screening mammography in this age group should be performed only if a patient expresses interest, mainly due to the non-significant mortality rate reduction and high overdiagnosis and overtreatment rates (Oeffinger, Fontham, & Etzioni, 2015) (Nelson, et al., 2016) (Barbeau, et al., 2017). We know a few risk factors that can apply to this age group (early menarche, late menopausal onset, hormone therapy, and family history of postmenopausal breast cancer). Another example is estrogen-positive breast cancer, whose incidence rates have been increasing in ages 40 to 49 and have not changed in older groups (Desantis, Howlader, Cronin, & Jemal, 2011). While there may be other factors that would determine the need for mammography in certain patients in this age group, there is no established risk model for this purpose.

Moreover, screening mammography is intended to determine the current risk of breast cancer without guiding future risks. In this work, we want to address this knowledge gap by developing a model that can be implemented when a female patient aged 40 to 70 attends a mammography screening to predict her 3-year future risk. This model would complement current screening where the patient would not only know whether they have current breast cancer, but also be able to quantify her future risk.

# Objective

The objective is to develop and validate patient-level prediction models for the following prediction problems:

| Target Cohorts | Outcome Cohorts | Time at Risk |
| --- | --- | --- |
| [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms | [WoO 2019] O: Breast cancer (2 occurrences) | [Time at Risk Settings #1] Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 1095, Add Exposure Days to End: FALSE |
| [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer | [WoO 2019] O: Breast cancer (2 occurrences) | [Time at Risk Settings #1] Risk Window Start: 1, Add Exposure Days to Start: FALSE, Risk Window End: 1095, Add Exposure Days to End: FALSE |

# Methods

## Study Design

This study will follow a retrospective, observational, patient-level prediction design. We define 'retrospective' to mean the study will be conducted using data already collected prior to the start of the study. We define 'observational' to mean there is no intervention or treatment assignment imposed by the study. We define 'patient-level prediction' as a modeling process wherein an outcome is predicted within a time at risk relative to the target cohort start and/or end date. Prediction is performed using a set of covariates derived using data prior to the start of the target cohort.

Figure 1, illustrates the prediction problem we will address. Among a population at risk, we aim to predict which patients at a defined moment in time (t = 0) will experience some outcome during a time-at-risk. Prediction is done using only information about the patients in an observation window prior to that moment in time.

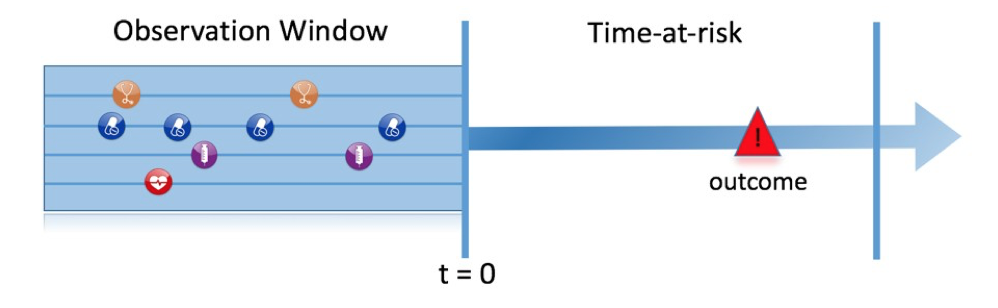


Figure 1: The prediction problem

We follow the PROGRESS best practice recommendations for model development and the TRIPOD guidance for transparent reporting of the model results (Steyerberg, et al., 2013)(Collins, 2017).

## Data Source(s)

Data owners from the OHDSI Research Network are invited to participate in this protocol. The following data sources have committed to complete this analysis for a presentation at the 2019 OHDSI US Symposium:

|  |  |  |
| --- | --- | --- |
| **Data Source Name** | **Contributor** | **Description** |
| Optum© de-identified Electronic Health Record Dataset (EHR) | Janssen | Optum’s de-identified Electronic Health Record data a medical records database. The medical record data includes clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using Natural Language Processing (NLP). |
| IBM MarketScan® Commercial Database (CCAE) | Janssen | Data from individuals enrolled in United States employer-sponsored insurance health plans. The data includes adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. Additionally, it captures laboratory tests for a subset of the covered lives. The patients in this database are aged under 65. |
| IBM MarketScan® Multi-State Medicaid Database (MDCD) | Janssen | Adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. The dataset lacks lab result data. |
| IBM MarketScan® Medicare Supplemental Database (MDCR) | Janssen | Represents health services of retirees (aged 65 or older) in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. These data include adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy). Additionally, it captures laboratory tests for a subset of the covered lives. |
| IQVIA Hospital Charge Detail Masters (CDM) | IQVIA | Anonymized patient level data are sourced from hospital charge detail masters (CDM) and collected from resource management software within short-term, acute-care and non-federal hospitals. |
| IQVIA LRxDx OpenClaims | IQVIA | Pre-adjudicated claims at the anonymized patient level collected from office-based physicians and specialists via office management software and clearinghouse switch sources for the purpose of reimbursement. We do have adjudicated claims for a subset of the medical claims data. |
| Stanford Healthcare STARR | Stanford | Electronic health record data derived from all patients treated as outpatients and inpatients at Stanford Hospital and Clinics from 1995 to 2019, including structured clinical data and unstructured clinical notes. |

Additional data sources may be added as this analysis continues to reiterate. This protocol will be updated to reflect additions to this list. All data are maintained in an OMOP compliant database (OHDSI, 2019). No data are pooled across sites – meaning each site individually runs the study package on their data and only transmits results. This OHDSI study package is published for use on the OHDSI GitHub: <https://github.com/OHDSI/StudyProtocols/tree/master/finalWoo>

## Study Populations

### Target Cohort(s) [T]

**Target #1 – Cohort Name in ATLAS:** [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms

**Initial Event Cohort**

People having any of the following:

* a procedure of Screening mammography 4
  + with age between 40 and 74 (inclusive)
  + gender is any of: FEMALE

with continuous observation of at least 1095 days prior and 0 days after event index date, and limit initial events to: **all events per person.**

**Inclusion Rules**

*Inclusion Criteria #1: No mammography in prior 700 days*

Having all of the following criteria:

* exactly 0 occurrences of a procedure of Screening mammography 4

where event starts between 700 days Before and 1 days Before index start date

* and exactly 0 occurrences of an observation of Screening mammography 4

where event starts between 700 days Before and 1 days Before index start date

*Inclusion Criteria #2: No neoplasm disease prior or up to 90 days following*

Having all of the following criteria:

* exactly 0 occurrences of a condition occurrence of Broad Malignancy1

where event starts between all days Before and 90 days After index start date

*Inclusion Criteria #3: No history of breast cancer all days prior and 90 days after*

Having all of the following criteria:

* exactly 0 occurrences of an observation of History of malignant neoplasm of breast2

where event starts between all days Before and 90 days After index start date

*Inclusion Criteria #4: No prior hormone antagonists 21603829*

Having all of the following criteria:

* exactly 0 occurrences of a drug exposure of hormone antagonists3

where event starts between all days Before and 0 days After index start date

Limit qualifying cohort to: **all events per person.**

**End Date Strategy**

*Date Offset Exit Criteria*

This cohort definition end date will be the index event's start date plus 0 days

**Target #2 – Cohort Name in ATLAS:** [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer

**Initial Event Cohort**

People having any of the following:

* a procedure of Screening mammography 4
  + with age between 40 and 74 (inclusive)
  + gender is any of: FEMALE

with continuous observation of at least 1095 days prior and 0 days after event index date, and limit initial events to: **all events per person.**

**Inclusion Rules**

*Inclusion Criteria #1: No mammography in prior 700 days*

Having all of the following criteria:

* exactly 0 occurrences of a procedure of Screening mammography 4

where event starts between 700 days Before and 1 days Before index start date

* and exactly 0 occurrences of an observation of Screening mammography 4

where event starts between 700 days Before and 1 days Before index start date

*Inclusion Criteria #2: No breast cancer prior or up to 90 days following*

Having all of the following criteria:

* exactly 0 occurrences of a condition occurrence of Breast cancer1

where event starts between all days Before and 90 days After index start date

* and exactly 0 occurrences of an observation of Breast cancer1

where event starts between all days Before and 90 days After index start date

*Inclusion Criteria #3: No history of breast cancer all time prior and 90 days after*

Having all of the following criteria:

* exactly 0 occurrences of an observation of History of malignant neoplasm of breast2

where event starts between all days Before and 90 days After index start date

*Inclusion Criteria #4: no prior hormone* *antagonists*

Having all of the following criteria:

* exactly 0 occurrences of a drug exposure of hormone antagonists 3

where event starts between all days Before and 0 days After index start date

Limit qualifying cohort to: **all events per person.**

**End Date Strategy**

*Date Offset Exit Criteria*

This cohort definition end date will be the index event's start date plus 0 days

| Cohort ID | Cohort Name | Description |
| --- | --- | --- |
| 10631 | [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms | Women with between the ages of 40-74 with a first mammogram with at least 2 years of time pre-index with no prior malignant neoplasms (all kinds) |
| 10845 | [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer | Women with between the ages of 40-74 with a first mammogram with at least 2 years of time pre-index with no prior breast cancer |

### Outcome Cohorts(s) [O]

**Outcome #1 – Cohort Name in ATLAS:** [WoO 2019] O: Breast cancer (2 occurrences)

**Initial Event Cohort**

People having any of the following:

* a condition occurrence of Primary malignant neoplasm of breast1

Having all of the following criteria:

* + - at least 1 occurrences of a condition occurrence of Primary malignant neoplasm of breast1

where event starts between 1 days After and all days After index start date

with continuous observation of at least 0 days prior and 0 days after event index date, and limit initial events to: **all events per person.**

Limit qualifying cohort to: **all events per person.**

**End Date Strategy**

*Date Offset Exit Criteria*

This cohort definition end date will be the index event's start date plus 0 days

**Cohort Collapse Strategy:**

Collapse cohort by era with a gap size of 0 days.

| Cohort ID | Cohort Name | Description |
| --- | --- | --- |
| 10082 | [WoO 2019] O: Breast cancer (2 occurrences) | People with a documented occurrence of breast cancer and another occurrence of breast cancer at any point in time after (2 occurrences or more) |

### Time at Risk

| Time at Risk |
| --- |
| [Time at Risk Settings #1]  Risk Window Start: 1 day after index,  Add Exposure Days to Start: FALSE (no days added) Risk Window End: 1095 after index  Add Exposure Days to End: FALSE (no days added) |

### Additional Population Settings

***Population Settings #1***

| Item | Settings |
| --- | --- |
| minTimeAtRisk | 364 |
| requireTimeAtRisk | FALSE |
| addExposureDaysToStart | FALSE |
| riskWindowStart | 1 |
| washoutPeriod | 1095 |
| addExposureDaysToEnd | FALSE |
| includeAllOutcomes | TRUE |
| priorOutcomeLookback | 99999 |
| binary | TRUE |
| removeSubjectsWithPriorOutcome | TRUE |
| riskWindowEnd | 1095 |
| firstExposureOnly | FALSE |

## Statistical Analysis Method(s)

### Algorithms

| Algorithm | Description |
| --- | --- |
| Lasso Logistic Regression | Lasso logistic regression belongs to the family of generalized linear models, where a linear combination of the variables is learned and finally a logistic function maps the linear combination to a value between 0 and 1. The lasso regularization adds a cost based on model complexity to the objective function when training the model. This cost is the sum of the absolute values of the linear combination of the coefficients. The model automatically performs feature selection by minimizing this cost. We use the Cyclic coordinate descent for logistic, Poisson and survival analysis (Cyclops) package to perform large-scale regularized logistic regression: https://github.com/OHDSI/Cyclops |

### Model Evaluation

The following evaluations will be performed on the model:

| Evaluation | Description |
| --- | --- |
| Box Plots | The prediction distribution boxplots are box plots for the predicted risks of the people in the test set with the outcome (class 1: blue) and without the outcome (class 0: red). |
| Calibration Plot | The calibration plot shows how close the predicted risk is to the observed risk. The diagonal dashed line thus indicates a perfectly calibrated model. The ten (or fewer) dots represent the mean predicted values for each quantile plotted against the observed fraction of people in that quantile who had the outcome (observed fraction). The straight black line is the linear regression using these 10 plotted quantile mean predicted vs observed fraction points. The two blue straight lines represented the 95% lower and upper confidence intervals of the slope of the fitted line. |
| Demographic Summary Plot | This plot shows for females and males the expected and observed risk in different age groups together with a confidence area. |
| Precision Recall Plot | The precision-recall curve is valuable for dataset with a high imbalance between the size of the positive and negative class. It shows the tradeoff between precision and recall for different threshold. High precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall). A high area under the curve represents both high recall and high precision. |
| Prediction Distribution Plots | The preference distribution plots are the preference score distributions corresponding to i) people in the test set with the outcome (red) and ii) people in the test set without the outcome (blue). |
| ROC Plot | The ROC plot plots the sensitivity against 1-specificity on the test set. The plot shows how well the model is able to discriminate between the people with the outcome and those without. The dashed diagonal line is the performance of a model that randomly assigns predictions. The higher the area under the ROC plot the better the discrimination of the model. |
| Smooth Calibration Plot | Similar to the traditional calibration shown above the Smooth Calibration plot shows the relationship between predicted and observed risk. the major difference is that the smooth fit allows for a more fine grained examination of this. Whereas the traditional plot will be heavily influenced by the areas with the highest density of data the smooth plot will provide the same information for this region as well as a more accurate interpretation of areas with lower density. the plot also contains information on the distribution of the outcomes relative to predicted risk. However the increased information game comes at a computational cost. It is recommended to use the traditional plot for examination and then to produce the smooth plot for final versions. |
| Test-Train Similarity Plot | The test-train similarity is presented by plotting the mean covariate values in the train set against those in the test set for people with and without the outcome. |
| Variable Scatter Plot | The variable scatter plot shows the mean covariate value for the people with the outcome against the mean covariate value for the people without the outcome. The size and color of the dots correspond to the importance of the covariates in the trained model (size of beta) and its direction (sign of beta with green meaning positive and red meaning negative), respectively. |

## Quality Control

The PatientLevelPrediction package itself, as well as other OHDSI packages on which PatientLevelPrediction depends, use unit tests for validation (R Core Team, 2013).

## Tools

This study is designed using OHDSI tools and run with R (OHDSI, 2019). More information about the tools can be found in the Appendix 'Study Generation Version Information'.

# Diagnostics

Reviewing the incidence rates of the outcomes in the target population prior to performing the analysis will allow us to assess its feasibility. The full table will be created during analysis and populated in a Shiny Application published to data.ohdsi.org.

Additionally, reviewing the characteristics of the cohorts provides insight into the cohorts being reviewed. The full table created during analysis and populated in a Shiny Application published to data.ohdsi.org.

# Data Analysis Plan

## Algorithm Settings

***Model Settings Settings #1 - LassoLogisticRegressionSettings***

| Covariates | Settings |
| --- | --- |
| seed |  |
| variance | 0.01 |

## Covariate Settings

The covariates (constructed using records on or prior to the target cohort start date) are used within this prediction mode include the following. Each covariate needs to contain at least 5e-04 subjects to be considered for the model.

***Covariate Settings #1***

| Covariates | Settings |
| --- | --- |
| VisitCountMediumTerm | FALSE |
| ObservationShortTerm | FALSE |
| shortTermStartDays | -30 |
| MeasurementRangeGroupShortTerm | FALSE |
| ConditionOccurrenceLongTerm | FALSE |
| DrugEraStartLongTerm | FALSE |
| VisitCountShortTerm | FALSE |
| Chads2Vasc | FALSE |
| ConditionGroupEraStartLongTerm | FALSE |
| ConditionEraShortTerm | FALSE |
| Dcsi | FALSE |
| DrugGroupEraLongTerm | TRUE |
| DrugGroupEraShortTerm | FALSE |
| ConditionEraStartLongTerm | FALSE |
| temporal | FALSE |
| DemographicsIndexMonth | FALSE |
| ConditionOccurrencePrimaryInpatientLongTerm | FALSE |
| ConditionEraAnyTimePrior | FALSE |
| addDescendantsToInclude | FALSE |
| ConditionGroupEraStartMediumTerm | FALSE |
| ProcedureOccurrenceLongTerm | FALSE |
| DrugExposureLongTerm | FALSE |
| DrugEraStartShortTerm | FALSE |
| DistinctIngredientCountMediumTerm | FALSE |
| DistinctMeasurementCountShortTerm | FALSE |
| MeasurementRangeGroupLongTerm | FALSE |
| ConditionGroupEraOverlapping | FALSE |
| MeasurementRangeGroupMediumTerm | FALSE |
| DrugGroupEraStartMediumTerm | FALSE |
| MeasurementAnyTimePrior | FALSE |
| MeasurementMediumTerm | FALSE |
| includedCovariateIds |  |
| ConditionOccurrenceAnyTimePrior | FALSE |
| DistinctConditionCountLongTerm | FALSE |
| MeasurementValueLongTerm | FALSE |
| DrugEraShortTerm | FALSE |
| DrugGroupEraAnyTimePrior | TRUE |
| DrugEraOverlapping | FALSE |
| ConditionOccurrencePrimaryInpatientAnyTimePrior | FALSE |
| ConditionEraMediumTerm | FALSE |
| ConditionEraOverlapping | FALSE |
| ConditionEraStartShortTerm | FALSE |
| ObservationAnyTimePrior | FALSE |
| VisitConceptCountShortTerm | FALSE |
| DemographicsEthnicity | FALSE |
| DistinctIngredientCountLongTerm | FALSE |
| ConditionOccurrencePrimaryInpatientShortTerm | FALSE |
| DemographicsAgeGroup | TRUE |
| DistinctProcedureCountShortTerm | FALSE |
| DistinctObservationCountMediumTerm | FALSE |
| includedCovariateConceptIds |  |
| DrugGroupEraStartShortTerm | FALSE |
| addDescendantsToExclude | FALSE |
| DrugEraLongTerm | FALSE |
| DistinctConditionCountShortTerm | FALSE |
| ConditionGroupEraShortTerm | FALSE |
| ConditionEraStartMediumTerm | FALSE |
| VisitCountLongTerm | FALSE |
| DemographicsRace | FALSE |
| ProcedureOccurrenceAnyTimePrior | FALSE |
| DistinctObservationCountLongTerm | FALSE |
| ProcedureOccurrenceMediumTerm | FALSE |
| CharlsonIndex | FALSE |
| DemographicsPriorObservationTime | FALSE |
| MeasurementShortTerm | FALSE |
| DistinctProcedureCountMediumTerm | FALSE |
| ConditionEraLongTerm | FALSE |
| DrugGroupEraStartLongTerm | FALSE |
| DemographicsGender | TRUE |
| DeviceExposureAnyTimePrior | FALSE |
| ObservationLongTerm | FALSE |
| DemographicsIndexYearMonth | FALSE |
| ConditionOccurrenceMediumTerm | FALSE |
| longTermStartDays | -365 |
| DemographicsAge | FALSE |
| DrugGroupEraOverlapping | FALSE |
| DistinctMeasurementCountLongTerm | FALSE |
| MeasurementRangeGroupAnyTimePrior | FALSE |
| DistinctConditionCountMediumTerm | FALSE |
| DrugGroupEraMediumTerm | FALSE |
| ProcedureOccurrenceShortTerm | FALSE |
| ObservationMediumTerm | FALSE |
| ConditionGroupEraAnyTimePrior | TRUE |
| Chads2 | FALSE |
| DrugExposureAnyTimePrior | FALSE |
| DeviceExposureLongTerm | FALSE |
| DemographicsTimeInCohort | FALSE |
| DistinctMeasurementCountMediumTerm | FALSE |
| MeasurementValueShortTerm | FALSE |
| DeviceExposureMediumTerm | FALSE |
| ConditionGroupEraStartShortTerm | FALSE |
| ConditionOccurrencePrimaryInpatientMediumTerm | FALSE |
| MeasurementLongTerm | FALSE |
| DemographicsIndexYear | FALSE |
| MeasurementValueMediumTerm | FALSE |
| DrugEraStartMediumTerm | FALSE |
| MeasurementValueAnyTimePrior | FALSE |
| DistinctObservationCountShortTerm | FALSE |
| DrugEraMediumTerm | FALSE |
| ConditionGroupEraLongTerm | TRUE |
| DrugExposureShortTerm | FALSE |
| DistinctIngredientCountShortTerm | FALSE |
| DeviceExposureShortTerm | FALSE |
| mediumTermStartDays | -180 |
| DemographicsPostObservationTime | FALSE |
| VisitConceptCountLongTerm | FALSE |
| VisitConceptCountMediumTerm | FALSE |
| excludedCovariateConceptIds |  |
| ConditionGroupEraMediumTerm | FALSE |
| DrugExposureMediumTerm | FALSE |
| DistinctProcedureCountLongTerm | FALSE |
| DrugEraAnyTimePrior | FALSE |
| endDays | 0 |
| ConditionOccurrenceShortTerm | FALSE |

## Model Development & Evaluation

To build and internally validate the models, we will partition the labelled data into a train set (75%) and a test set (25%).

The hyper-parameters for the models will be assessed using 3-fold cross validation on the train set and a final model will be trained using the full train set and optimal hyper-parameters.

The internal validity of the models will be assessed on the test set. We will use the area under the receiver operating characteristic curve (AUC) to evaluate the discriminative performance of the models and plot the predicted risk against the observed fraction to visualize the calibration. See 'Model Evaluation' section for more detailed information about additional model evaluation metrics.

## Analysis Execution Settings

There are 2 target cohorts evaluated for 1 outcomes over 1 models over 1 covariates settings and over 1 population settings. In total there are 2 analysis performed. For a full list refer to appendix 'Complete Analysis List'.

# Strengths & Limitations

**Strengths:**

* Collaboration – we are developing models across a number of data sites to gain insight into the models and this will also enable us to do large scale validation
* PLP framework – the prediction framework has been tested and includes state of the art methods
* Evaluation – we follow best practices for plots/evaluation metrics

**Limitations:**

* Missing data – some patients may have conditions that are not recorded in the databases. We treat these the same as the patient not having the condition.
* Misclassification – our phenotype for breast cancer may not identify all breast cancer patients or may incorrectly identify people without breast cancer.
* Lack of breast cancer predictors – many know risk factors such as smoking or family history of cancer may not be recorded into the datasets we use in this study

# Protection of Human Subjects

Confidentiality of patient records will be maintained always. All study reports will contain aggregate data only and will not identify individual patients or physicians. At no time during the study will the sponsor receive patient identifying information except when it is required by regulations in case of reporting adverse events.

# Plans for Disseminating & Communicating Study Results

Our main aim is to present the model and finding at the OHDSI 2019 symposium. We also plan to write a journal paper and disseminate the results using data.ohdsi.org as a shiny app, see <http://data.ohdsi.org/plpLive18Study/> as an example.  *here. >>*

# Appendices

## Study Generation Version Information

Skeleton Version: PatientLevelPredictionStudy - v0.0.1

Identifier / Organization: Janssen Research and Development

## Code List

***Concept Set #1 - Primary malignant neoplasm of breast***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | | CONCEPT\_CODE | | CONCEPT\_NAME | | | DOMAIN\_ID | | STANDARD\_CONCEPT\_CAPTION | | CONCEPT\_CLASS\_ID | | INVALID\_REASON | | isExcluded | | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | | 140960 | | 94297009 | | Secondary malignant neoplasm of female breast | Condition | | Standard | | Clinical Finding | | V | | TRUE | | TRUE | |
| Valid | | 81251 | | 126926005 | | Neoplasm of breast | Condition | | Standard | | Clinical Finding | | V | | FALSE | | TRUE | |
| Valid | | 72576 | | 269485000 | | Benign tumor of breast | Condition | | Standard | | Clinical Finding | | V | | TRUE | | TRUE | |
| Valid | | 81250 | | 189336000 | | Carcinoma in situ of breast | Condition | | Standard | | Clinical Finding | | V | | TRUE | | TRUE | |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 10082 | [WoO 2019] O: Breast cancer (2 occurrences) |

***Concept Set #2 - Screening mammography***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 4077697 | 24623002 | Screening mammography | Procedure | Standard | Procedure | V | FALSE | TRUE |
| Valid | 42627987 | 77067 | Screening mammography, bilateral (2-view study of each breast), including computer-aided detection (CAD) when performed | Procedure | Standard | CPT4 | V | FALSE | TRUE |
| Invalid | 2617289 | G0202 | Screening mammography, bilateral (2-view study of each breast), including computer-aided detection (cad) when performed | Procedure | Non-Standard | HCPCS | D | FALSE | TRUE |
| Valid | 4147961 | 268547008 | Screening for malignant neoplasm of breast | Procedure | Standard | Procedure | V | TRUE | TRUE |
| Valid | 46257421 | 77063 | Screening digital breast tomosynthesis, bilateral (List separately in addition to code for primary procedure) | Procedure | Standard | CPT4 | V | FALSE | TRUE |
| Valid | 2211809 | 77052 | Computer-aided detection (computer algorithm analysis of digital image data for lesion detection) with further review for interpretation, with or without digitization of film radiographic images; screening mammography (List separately in addition to code | Procedure | Standard | CPT4 | V | FALSE | TRUE |
| Valid | 2211814 | 77057 | Screening mammography, bilateral (2-view study of each breast) | Procedure | Standard | CPT4 | V | FALSE | TRUE |
| Valid | 2106158 | 3014F | Screening mammography results documented and reviewed (PV) | Observation | Standard | CPT4 | V | FALSE | TRUE |
| Valid | 46257521 | 77062 | Digital breast tomosynthesis; bilateral | Procedure | Standard | CPT4 | V | FALSE | TRUE |
| Valid | 46257687 | 77061 | Digital breast tomosynthesis; unilateral | Procedure | Standard | CPT4 | V | FALSE | TRUE |
| Valid | 43531505 | 609223006 | MRI of breast for screening for malignant neoplasm | Procedure | Standard | Procedure | V | TRUE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 10631 | [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms |
| 10845 | [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer |

***Concept Set #3 - Broad Malignancy***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 433435 | 109355002 | Carcinoma in situ | Condition | Standard | Clinical Finding | V | FALSE | TRUE |
| Valid | 443392 | 363346000 | Malignant neoplastic disease | Condition | Standard | Clinical Finding | V | FALSE | TRUE |
| Valid | 438112 | 55342001 | Neoplastic disease | Condition | Standard | Clinical Finding | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 10631 | [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms |

***Concept Set #4 - History of malignant neoplasm of breast***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 4324190 | 429087003 | History of malignant neoplasm of breast | Observation | Standard | Context-dependent | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 10631 | [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms |
| 10845 | [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer |

***Concept Set #5 - hormone antagonists***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 21603829 | L02B | HORMONE ANTAGONISTS AND RELATED AGENTS | Drug | Classification | ATC 3rd | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 10631 | [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms |
| 10845 | [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer |

***Concept Set #6 - Breast cancer***

| INVALID\_REASON\_CAPTION | CONCEPT\_ID | CONCEPT\_CODE | CONCEPT\_NAME | DOMAIN\_ID | STANDARD\_CONCEPT\_CAPTION | CONCEPT\_CLASS\_ID | INVALID\_REASON | isExcluded | includeDescendants |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Valid | 37103688 | 10006289 | Benign and malignant breast neoplasms | Condition | Classification | HLT | V | FALSE | TRUE |
| Valid | 4324190 | 429087003 | History of malignant neoplasm of breast | Observation | Standard | Context-dependent | V | FALSE | TRUE |
| Valid | 4111018 | 255058005 | Neoplasm of thorax | Condition | Standard | Clinical Finding | V | FALSE | TRUE |

Cohorts that use this Concept Set:

| Cohort ID | Cohort Name |
| --- | --- |
| 10845 | [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer |

## Complete Analysis List

Below is a complete list of analysis that will be performed. Definitions for the column 'Covariate Settings ID' can be found above in the 'Covariate Settings' section. Definitions for the 'Population Settings Id' can be found above in the 'Additional Population Settings' section.

| ID | Target Cohort Name | Outcome Cohort Name | Model Settings Id | Model Settings Description | Covariate Settings ID | Population Settings ID |
| --- | --- | --- | --- | --- | --- | --- |
| 1 | [WoO 2019] T1 - first mammography in 2 years and no prior neoplasms | [WoO 2019] O: Breast cancer (2 occurrences) | 1 | Lasso Logistic Regression | 1 | 1 |
| 2 | [WoO 2019] T2 - first mammography in 2 years and no prior breast cancer | [WoO 2019] O: Breast cancer (2 occurrences) | 1 | Lasso Logistic Regression | 1 | 1 |

*< add mod here >>*

# References

Barbeau, P., Stevens, A., Beck, A., Skidmore, B., Arnaout, A., & Brackstone, M. (2017). Retrieved from Breast Cancer Screening: Part A. An Evidence Report to Inform an Update of the Canadian Task Force on Preventive Health Care 2011 Guidelines: https://canadiantaskforce.ca/wp-content/uploads/2018/11/Evidence-Report-Breast-Cancer-Screening\_FINAL.pdf

Collins, G. e. (2017, 02 01). *'Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement*. Retrieved from https://www.equator-network.org/reporting-guidelines/tripod-statement/

Desantis, C., Howlader, N., Cronin, K., & Jemal, A. (2011). Breast cancer incidence rates in US women are no longer declining. *Cancer Epidemiol Biomarker Prev, 20*(5), 733-739.

Howlader, N., Noone, A., & al, K. M. (2019). *SEER Cancer Statistics Review 1975-2016: Table 1.11. Median age of cancer patients at diagnosis, 2012-2016.* Bethesda, MD: National Cancer Institute.

Nelson, H., Fu, R., Cantor, A., Pappas, M., Daeges, M., & Humphrey, L. (2016, January 12). Effectiveness of Breast Cancer Screening: Systematic Review and Meta-analysis to Update the 2009 U.S. Preventive Services Task Force Recommendation. *Ann Intern Med*(164), 244-255. doi:10.7326/M15-0969

Oeffinger, K., Fontham, E., & Etzioni, R. e. (2015). Breast Cancer Screening for Women at Average Risk: 2015 Guideline Update From the American Cancer Society. *JAMA*(314), 1599-1614. doi:10.1001/jama.2015.12783

OHDSI. (2019, 08 09). *OMOP Common Data Model*. Retrieved from OMOP Common Data Model (CDM): https://github.com/OHDSI/CommonDataModel

Qaseem A, Lin JS, Mustafa RA, Horwitch CA, Wilt TJ, for the Clinical Guidelines Committee of the American College of Physicians. (2019, April 09). Screening for Breast Cancer in Average-Risk Women: A Guidance Statement From the American College of Physicians. *Ann Intern Med*(170), 547-560. doi:10.7326/M18-2147

R Core Team. (2013). *R: A language and environment for statistical computing*. Retrieved from R Foundation for Statistical Computing: http://www.R-project.org/

Steyerberg, E., Moons, K., van der Windt, D., Hayden, J., Perel, P., Schroter, S., . . . Altman, D. (2013, 10). Prognosis Resaerch Study (PROGRESS) 3: Prognostic model research. *PLoS Medicine*(e1001381), 2. doi:10.1371/journal.pmed.1001381