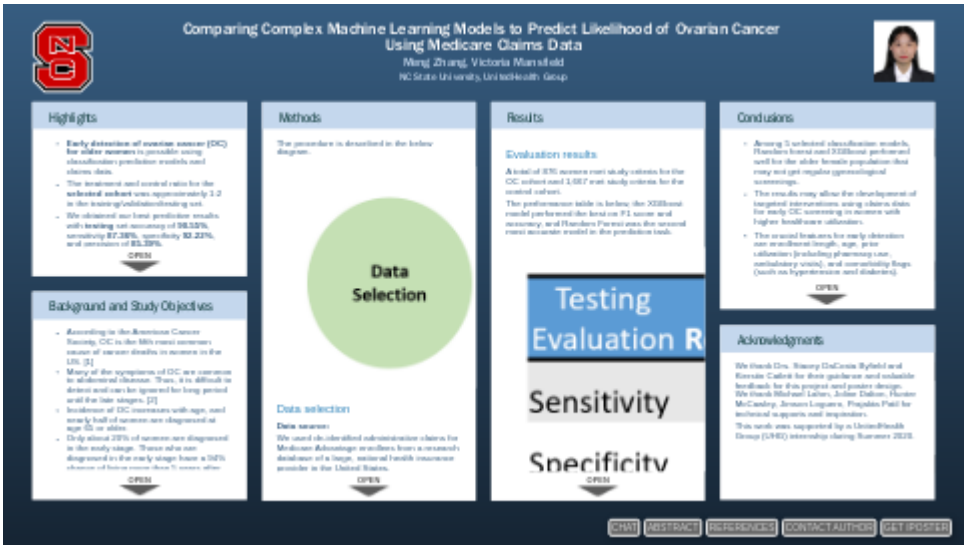
Comparing Complex Machine Learning Models to Predict Likelihood of Ovarian Cancer Using Medicare Claims Data



Meng Zhang, Victoria Mansfield

NC State University, UnitedHealth Group

PRESENTED AT:

HIGHLIGHTS

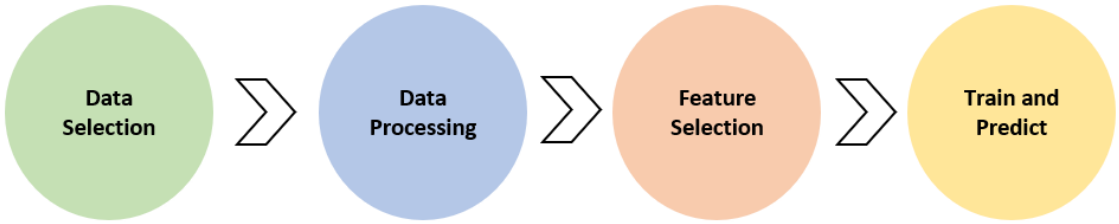
- **Early detection of ovarian cancer (OC) for older women** is possible using classification predictive models and claims data.
- The treatment and control ratio for the **selected cohort** was approximately 1:2 in the training/validation/testing set.
- We obtained our best predictive results with **testing** set accuracy of **90.55%**, sensitivity **87.36%**, specificity **92.22%**, and precision of **85.39%**.
- The visualization of the best model is presented by **SHAP values** in the Results section, which describes the impacts on model prediction for each feature.

BACKGROUND AND STUDY OBJECTIVES

- According to the American Cancer Society, OC is the fifth most common cause of cancer deaths in women in the US. [1]
- Many of the symptoms of OC are common to abdominal disease. Thus, it is difficult to detect and can be ignored for long period until the late stages. [2]
- Incidence of OC increases with age, and nearly half of women are diagnosed at age 65 or older.
- Only about 20% of women are diagnosed in the early stage. Those who are diagnosed in the early stage have a 94% chance of living more than 5 years after diagnosis. [1]
- The average cost of care for OC in the first year is ≈ $100,000. We can decrease the costs if the cancer is detected early. [3]
- This study aims to compare several models for predicting OC and identify healthcare utilization patterns, which may allow for intervention and reduce time to diagnosis.

METHODS

The procedure is described in the below diagram.



Data selection

Data source:

We used de-identified administrative claims for Medicare Advantage enrollees from a research database of a large, national health insurance provider in the United States.

The treatment cohort (with OC diagnosis) was selected first, and propensity score matching (PSM) was used to match the control cohort (without OC diagnosis). The treatment and control ratio is 1:2.

Criteria for treatment group:

1. Woman age between 66 to 85,
2. Has continuous enrollment (CE) for ≥30 months prior to first OC claim,
3. OC was identified by 2+ claims with ICD-10 C56.x between 30 and 365 days apart from Jan. 2017- May 2020.

Criteria for control group:

1. Woman age between 66 to 85,
2. CE at ≥30 months prior to index month 0,[a]
3. Have 0 or 1 claim with ICD-10 C56.x within one year from Jan. 2017-May 2020.

[a] Index month 0 is at 12 months before the latest claim date in the record, i.e., if a member's last visit date in the claims history was May 2018, then the index month 0 was May 2017.

Feature generation:
In addition to demographic information (age, enrollment length, business line, etc.), the features generated from claims data can be described below.

The claims history is separated according to the number of years prior to diagnosis. For the first year prior to diagnosis, the information from the last 3 months is removed.

- Disease flag prior x year: the comorbidities flag (e.g., diabetes, hypertension) is flagged if a member is diagnosed with the disease in the prior x year (x is range from 1 to 4).
- Year summary statistics prior x year: including the summary of the number of inpatient/ emergency/ ambulatory/ pharmacy(Rx) visits, and total allowance amount.

Data Processing

1. If the created feature has a large proportion (>50%) of missing values, then the feature is excluded.
2. For the missing values in the disease flag, we impute with 0 to represent no disease reported in this time period.
3. The string-type data was transferred to integer-type data using a label encoder.

Feature Selection

Remove the feature that showed strong collinearity with others.

Data split:

Data was separated into 64% training, 16% validation, and 20% testing.

Train and Predict

We used 5 predictive algorithms: Lasso Logistic Regression, Decision Tree, Random Forest, CatBoost, and XGBoost.

The data was normalized to the range [0,1]. The training set is used to train the models and build initial model parameters (e.g., weights and biases). Model parameters were tuned using validation data by applying the grid search on hyperparameter values. Finally, the prediction results were evaluated by a testing set.

Evaluation Metrics:

Sensitivity = TP/(TP+FN), measures how often a test correctly generates a positive result for people who have the condition that is being tested for.

Precision (PPV) = TP/(TP+FP), the positive prediction rate measures how often a test correctly finds a positive result for all the examples predicted as positive.

Specificity = TN/(FP+TN), measures a test's ability to correctly generate a negative result for people who do not have the condition that is being tested for.

Accuracy = (TP+TN)/(TP+TN+FP+FN), the probability of classifying incorrect group.

F1 score = 2*PPV*Sensitivity/(PPV+Sensitivity), taking into account both sensitivity and precision.

AUC is a metric that evaluated the area bounded by true positive rate and false positive rate.

TP: true positive, TN: true negative, FP: false positive, FN: false negative.

RESULTS

Evaluation results

A total of 876 women met study criteria for the OC cohort and 1,667 met study criteria for the control cohort.

The performance table is below, the XGBoost model performed the best on F1 score and accuracy, and Random Forest was the second most accurate model in the prediction task.

Testing Evaluation	Logistic Regression	Decision Tree	Random Forest	CatBoost	XGBoost
Sensitivity	64.37%	86.21%	**95.59%**	34.66%	**87.36%**
Specificity	81.74%	89.82%	**88.17%**	100%	**92.22%**
Precision	64.74%	81.52%	**74.71%**	100%	**85.39%**
Accuracy	75.79%	88.58%	**90.16%**	35.43%	**90.55%**
F1 Score	64.55%	83.80%	**83.87%**	51.48%	**86.36%**
AUC	73.05%	88.01%	**91.88%**	67.33%	**89.79%**

Table 1. Prediction evaluation on 5 classification models for testing data.

Visualization of Impacting Factors

The important ranking of features was also investigated for the XGBoost model. We used SHAP values to show the positive or negative quantified impacts for the training features.

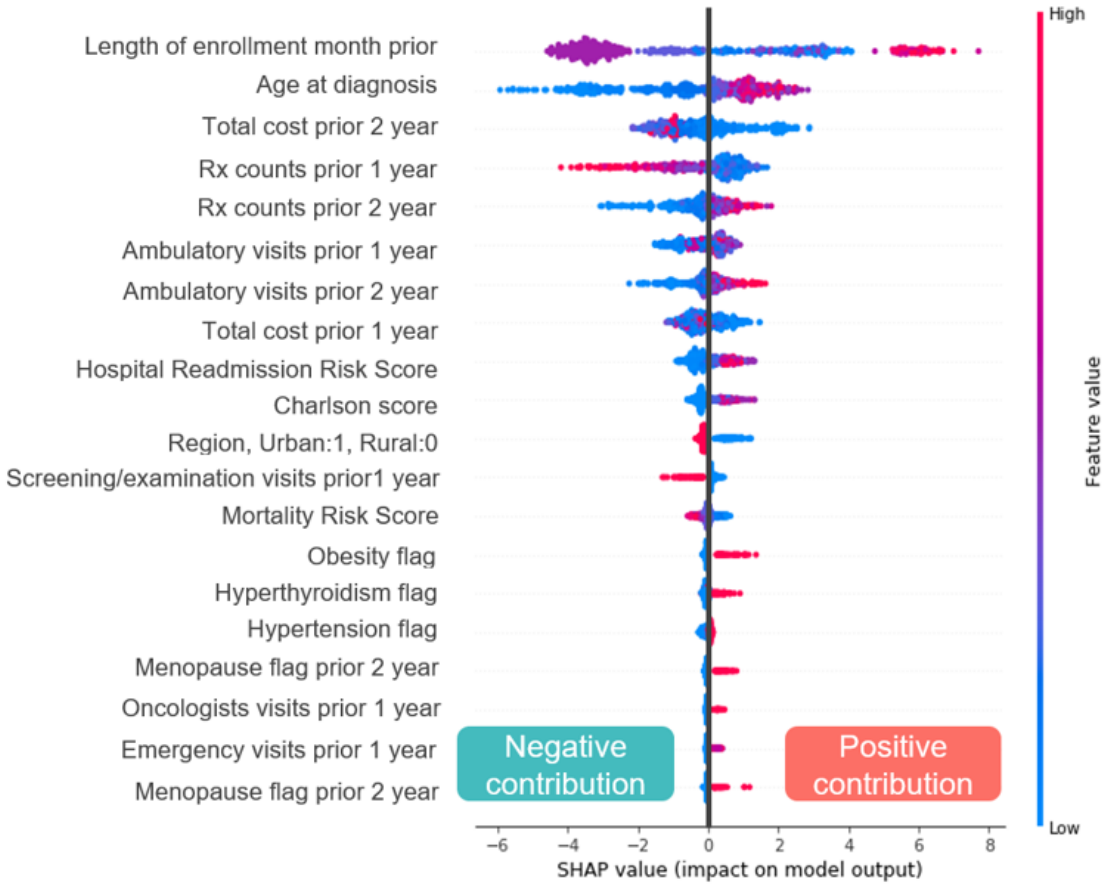


Figure 1. The SHAP values explaining the XGBoost result.

The SHAP values in Figure 1 indicate the following:

- Top important features included length of CE prior to diagnosis, age, and general utilization (Rx, ambulatory, and total allowance amounts).
- Longer enrollment and older age will result in a higher chance of diagnosis with OC.
- Rx use in the year prior to diagnosis was associated with lowered risk of OC, while Rx uses two years prior to the diagnosis was associated with increased risk.
- Menopause and comorbidities (such as obesity, hyperthyroidism, and hypertension) slightly increase the risk of OC.

CONCLUSIONS

- Among 5 selected classification models, Random forest and XGBoost performed well for the older female population that may not get regular gynecological screenings.
- The results may allow the development of targeted interventions using claims data for early OC screening in women with higher healthcare utilization.
- The crucial features for early detection are enrollment length, age, prior utilization (including pharmacy use, ambulatory visits), and comorbidity flags (such as hypertension and diabetes).
- Future work includes building more features specific to the pharmacy and ambulatory utilization and working towards identifying better candidate cohorts in the younger population.
- The ultimate goal is to develop predictive models that use claims data for earlier identification of women with OC from the general population.

ACKNOWLEDGMENTS

We thank Drs. Stacey DaCosta Byfield and Kierstin Catlett for their guidance and valuable feedback for this project and poster design. We thank Michael Lahm, Joline Dalton, Hunter McCawley, Jimson Loguero, Prajakta Patil for technical supports and inspiration.

This work was supported by a UnitedHealth Group (UHG) internship during Summer 2020.

ABSTRACT

Title: Comparing Complex Machine Learning Models to Predict Likelihood of Ovarian Cancer using Medicare Claims Data

Author(s): Zhang, M; Mansfield, V

NC State University, Raleigh, NC; UnitedHealth Group, Minnetonka, MN

Objectives:

Ovarian cancer (OC) is difficult to detect and largely goes undiagnosed until late stages. Incidence of OC increases with age, and nearly half of women are diagnosed at age 65 or greater. This study aims to compare several models for predicting ovarian cancer. The goal is to identify healthcare utilization patterns that are associated with an increased OC risk that may allow for intervention and reduce time to diagnosis.

Methods:

Using a de-identified claims database from 01/2017-05/2020, Medicare Advantage enrollees with OC were identified: 2+ claims with ICD-10 C56.x between 30 and 365 days apart. Patients were required to have continuous enrollment (CE) for ≥30 months prior to first OC claim and be between 66 to 85 years old at time of diagnosis. A control cohort was selected from women with equivalent age and CE criteria. We compared five algorithms in predicting likelihood of OC: LASSO logistic regression, decision tree, random forest, XGBoost, and CatBoost.

Results:

A total of 876 women met study criteria for the OC cohort and 1,667 met study criteria for the control cohort. Random forest (sensitivity 0.96, specificity 0.88, precision 0.75) and XGBoost (sensitivity 0.87, specificity 0.92, precision 0.85) were the best performing models. Length of continuous enrollment, age, and pharmacy utilization in the two years prior had the highest feature importance for both models. Ambulatory utilization and total payer costs were also identified as important features in random forest and XGBoost respectively.

Conclusions:

Random forest and XGBoost performed well at predicting OC in a higher-risk Medicare Advantage population and may allow development of targeted interventions using claims data for early OC screening in women with higher healthcare utilization. Future iterations will explore implications of model sensitivity and specificity; minimum enrollment criteria will be assessed, and analysis will be extended to investigate model performance in a younger commercial population.

REFERENCES

[1] Cancer Facts and Figures 2019. *Cancer.org*. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf (https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf). Accessed 13 April 2021.

[2] Stewart C, Ralyea C, Lockwood S, et al. Ovarian Cancer: An Integrated Review. *Seminars in Oncology Nursing*.2019;35(2):151-156. doi: 10.1016/j.soncn.2019.02.001.

[3] Bercow AS, Chen L, Chatterjee S, et al. Cost of Care for the Initial Management of Ovarian Cancer. *Obstet Gyneco*. 2017;130(6):1269-1275. doi:10.1097/AOG.0000000000002317