Breast Cancer High Drug Cost Member Predictive Model Creation and Independent Factor Identification

K. Thompson[1], P. Bryan, PharmD[1], Y. Qiu, MS[1], S. Champaloux, PhD, MPH[1], P.P. Gleason, PharmD[1,2]. [1]Prime Therapeutics LLC, Eagan, MN, United States; [2]University of Minnesota College of Pharmacy, Minneapolis, MN, United States.

BACKGROUND

- Breast cancer is the most common malignancy in women in the United States with average annual costs ranging from $50,000 with stage 0 to over $200,000 for stage IV, frequently driven by drug costs on both medical and pharmacy benefits.[1] High drug spend breast cancer member forecasting and associated predictive factors will provide opportunities to optimize cost-effective breast cancer therapy.
- One in eight women will be diagnosed with breast cancer in their lifetime. In 2021 alone, the number of new breast cancer cases in the U.S. reached 281,500.[2] Because of breast cancer's pervasive reach and its high-cost treatments, it ranks as a high spend condition within the oncology category.
- Recommended therapy for breast cancer is organized by cancer stage and, at the highest level, by subcategorization by the expression of hormone receptors (HR) and HER2 into: Luminal (HR-positive/HER2-negative), human epidermal growth factor receptor-2 (HER2) positive, and triple negative (HR-negative/HER2-negative) breast cancers.
- A significant amount of spending within breast cancer is drug spending, both on the medical benefit and the pharmacy benefit. This creates an opportunity for pharmacy benefit manager to provider cost effective drug therapy optimization considerations.
- Prime Therapeutics has successfully developed a predictive model to identify "drug super spenders" in the commercial population. In conjunction with clinical rules and pharmacist outreach, the drug super spender predictive model has demonstrated savings for health insurance providers.[3]
- To our knowledge, a breast cancer high drug cost predictive model to aid in ensuring cost effective drug therapy has not been developed.

OBJECTIVES

- Create a breast cancer total drug spend next 12-month predictive model utilizing integrated medical and pharmacy claims, and identify independent predictive factors associated with members having greater than $100,000 annual drug costs.

4085-B © Prime Therapeutics LLC 10/22
2900 Ames Crossing Road, Eagan, MN 55121

Academy of Managed Care Pharmacy (AMCP)
Nexus Meeting, Oct. 12, 2022, National Harbor, MD

PATRICK GLEASON, 800.858.0723, ext. 5190
pgleason@primetherapeutics.com

METHODS

- We reviewed medical and pharmacy claims from approximately 16 million commercial members from 2018 to 2020.
- We created a member with breast cancer (registry) based on four rules that use pharmacy claims, medical claims, and Optum Symmetry Episode Treatment Groups (ETG)[4] software to identify a breast cancer population in which a person must qualify based on one or more of the rules.

Breast Cancer Member Identification Rules:

1. Has at least one breast cancer diagnosis in the past year and a base ETG for breast cancer in the past year
2. Has one medical claim current procedural terminology (CPT) code for a mastectomy/lumpectomy in the past year and has a diagnosis of breast cancer on the same claim
3. Has at least one in situ breast cancer diagnosis in the past year and a base ETG for breast cancer in the past year
4. Is female and taking a high-cost drug primarily used for breast cancer through the pharmacy benefit in the past year

- Monthly data for training, validation and testing started with a simulated run date in September 2019 and ended in May 2020. The dates account for requirements of a look back period as well as a look forward period (i.e., a simulated run in September 2019 required data from September 1, 2018 to August 31, 2020). We chose to exclude six months of data, meeting the time constraint requirements due to typical lag of medical claims in processing time and file intake frequency.
- Approximately 2.3 million member records were identified for inclusion in the model training data. We sampled up to two simulated month runs in the time series data from each person for the model training data. Model was split 60% for training, 20% for validation and 20% for test data.
- Performance metrics were identified ahead of model builds based on the potential business use cases. For classification (predicting whether a member will have >= $100,000 total drug spend in the next 12 months), we selected area under the curve (AUC), log loss and F2. For regression (predicting amount of total drug spend in the next 12 months), we selected root mean squared error (RMSE). AUC and log loss were chosen for their overall model assessments. F2 was chosen so that there was greater emphasis on improving false negatives (people that become high spenders but are predicted to not be high spenders). RMSE was chosen over mean absolute error (MAE) to emphasize importance of outlier errors.
- In order for the model to progress to testing business use, we identified some benchmark models that it had to outperform including previous spend, always high spender, always not high spender, and others; see (Table 1)
- Initial exploratory data analysis showed a HER2 positive indicator and a metastases indicator having a strong positive relationship to total drug spend in the next 12 months.
- We explored multiple model types including random forest, extreme gradient boosting, gradient boosting machine, neural networks, linear regression, and others.

RESULTS

- In any given month, between 53,510 and 59,911 members during September 2019 to May 2020 were identified as having breast cancer, and between 5.1% to 5.4% had >$100,000 in annual drug spend.
- The test data showed a breast cancer identified member mean previous total drug spend in the prior 12 months of $13,775 and a median of $655.
- Initially an automated machine learning (AutoML) process, using the open-source H2O package in the R programming language led to identifying an XGBoost model for classification and a gradient boosting machine (GBM) model for the regression. We chose those models after further review. We allowed the results from the classification model and the classification using regression predictions to potentially differ.
- The models outperformed the benchmark for all performance metrics. RMSE on test data was $33,771 for the final model and $53,804 using the benchmark.
- Potential predictor variables were identified based on clinical background research, time series financial information, member demographics, financial risk scores, diagnosis codes, breast cancer type, previous spending, drug utilization, timing of surgery relative to drug therapy, and more. For clinical programs, the following variables were included: total previous drug spend in the last 12 months, use of drugs classifying the person as HER2 positive or negative (default) in the past year, whether they had breast cancer surgery in the past year, whether the Symmetry ETG classified a cancer with metastases in the past year, whether they had an HR positive diagnosis, whether they had an HR negative diagnosis, and whether they had an in situ breast cancer diagnosis in the past year.
- Key predictors identified included previous total drug spend in the past 12 months, use of drugs associated with HER2 positive status, metastases indicators, and hormone receptor positive or negative diagnoses.
- Key classification metrics of overall model performance show that the classification model outperformed the benchmark of using previous spending to predict future spending classification on unseen test data. Key classification metrics of threshold-dependent performance show the predictive model with a better F2 performance. The actual prevalence of high-cost members in test data was 5.15%; the classification model overestimates the prevalence of high-cost breast cancer members in exchange for better predictive power at identifying high-cost members. Key regression metrics show the regression model outperforms other model benchmarks significantly.
- Predictions were generated in production processes starting in December 2021. (See Tables 2, 3, and 4 for sample reporting from February 2022)

LIMITATIONS

- Pharmacy claims are updated in the data warehouse daily with almost no lag in time of receipt, whereas medical claims are updated in the data warehouse monthly and often lag one to three months in time of receipt from date of service. This limits our ability to predict as well as our business capabilities in determining a timely intervention process.
- Breast cancer stage is limited to medical coding accuracy and electronic medical health record data, and laboratory data were unavailable.
- The data used was limited to members and data in the commercial care line of business.
- Data used in 2020 could have been affected by the COVID-19 pandemic.
- Thoroughness of feature engineering, model identification, and tuning was attenuated in exchange for quicker feedback and response of business units.

TABLE 1

Performance Metrics: Predictive Models and Benchmarking on Test Data

Model or Benchmark Description	Model Type	AUC	Log loss	F2	Predicted Prevalence	RMSE
Classification model using the threshold maximizing F2	Machine learning prediction	0.9173	0.1089	0.6656	0.0812	N/A
Regression model	Machine learning prediction	N/A	N/A	N/A	N/A	$33,771
HER2 positive or metastases or >$100K previous spend	Benchmark	N/A	N/A	0.5981	0.0639	N/A
Previous spend	Benchmark	0.9099	0.2019	0.5216	0.0416	$53,804
HER2 positive or metastases	Benchmark	N/A	N/A	0.383	0.0462	N/A
HER2 positive	Benchmark	N/A	N/A	0.34	0.0414	N/A
Metastases	Benchmark	N/A	N/A	0.098	0.0070	N/A
Mean previous spend	Benchmark	N/A	N/A	N/A	N/A	$47,970
Median previous spend	Benchmark	N/A	N/A	N/A	N/A	$50,114

*N/A = not applicable

AUC = area under the receiver operating characteristics curve; higher values indicate better performance; possible values from 0 to 1. Log loss = logarithmic loss cost function; lower values indicate better performance; possible values from 0 to infinity. F2 = F score balancing precision and recall with recall weighted twice as much as precision; higher values indicate better performance; possible values from 0 to 1. Predictive prevalence = predicted proportion of population having >$100,000 drug spend. RMSE = root mean squared error; lower values indicate better performance; possible values from 0 to infinity.

HER2 = Members with HER2 positive indication based on drug use. Metastases = Members with metastatic cancer as identified by ETGs.

These findings demonstrate how the predictive models performed against other benchmark models. The classification model using the threshold maximizing F2 used the XGBoost algorithm and chose a threshold value that maximized the F2 score when a threshold was required for measuring classification performance. The regression model used the GBM algorithm. The previous spend benchmark model used the member's drug spending the last 12 months and predicted the same spending for the next 12 months. The benchmark model HER2 positive or metastases or >$100k previous spend predicted a member to be high spend if they had a HER2 positive indication, a metastases indication, or had >$100k in drug spending the past 12 months; the models for HER2 positive or metastases, HER2 positive, and metastases followed a similar prediction pattern. The mean previous spend benchmark model used the mean spend of all members the past 12 months and predicted every member to spend that amount in the next 12 months; median previous spend benchmark model followed a similar pattern.

TABLE 2

Sample Report – Inclusion and Enrollment with Rule Details, February 2022

Member	Rule 1 Inclusion	Rule 2 Inclusion	Rule 3 Inclusion	Rule 4 Inclusion	Line of Business
1	Yes	Yes	Yes	No	Fully Insured
2	Yes	No	No	Yes	HIM
3	Yes	No	No	No	HIM
4	Yes	No	No	No	Fully Insured
5	Yes	No	Yes	No	HIM
6	Yes	No	No	No	HIM
7	Yes	No	No	No	HIM
8	Yes	No	No	Yes	Fully Insured
9	No	No	Yes	No	HIM
10	No	No	No	Yes	Self Insured

*HIM = Health Insurance Marketplace.

This table describes how the members were included for reporting and predictions along with their current enrollment information. Breast cancer member identification rules: 1) Member has at least one breast cancer diagnosis in the past year and a base ETG for breast cancer in the past year. 2) Member has one medical claim current procedural terminology (CPT) code for a mastectomy/lumpectomy in the past year and has a diagnosis of breast cancer on the same claim. 3) Member has at least one in situ breast cancer diagnosis in the past year and a base ETG for breast cancer in the past year. 4) Member is female and taking a high-cost drug primarily used for breast cancer through the pharmacy benefit in the past year.

TABLE 3

Sample Report – Clinical Markers, February 2022

Member	Most Recent Breast Cancer Claim	Earliest Known Breast Cancer Diagnosis Claim	Most Recent Surgery Date	Most Recent In Situ Breast Cancer Diagnosis Claim	HER2 Positive	HR Positive	HR Negative	Metastases
1	01Oct2021	20Aug2020	23Apr2021	14Sep2021	No	Yes	No	No
2	28Jan2022	29Jul2021	Not found	Not found	No	Yes	No	No
3	15Dec2021	22Jul2020	Not found	Not found	No	Yes	No	No
4	30Jun2021	25Nov2020	Not found	Not found	No	Yes	No	No
5	16Dec2021	17Jan2020	Not found	20May2021	Yes	No	Yes	No
6	03Jan2022	06Jan2020	Not found	Not found	No	No	No	Yes
7	03Jan2022	07Jan2020	Not found	Not found	Yes	Yes	Yes	No
8	17Jan2022	10Sep2021	Not found	Not found	Yes	Yes	No	Yes
9	26Oct2021	11Jun2020	Not found	11Mar2021	No	No	No	No
10	25Jan2022	12Aug2021	Not found	Not found	No	No	No	No

*HER2 = human epidermal growth factor receptor 2, HR = hormone receptor.

This table describes known clinical information about the members. Sometimes diagnoses may not be found but does not guarantee a lack of a diagnosis due to possible causes of medical claim lag, incomplete historical claims data, or complete coding by medical providers.

TABLE 4

Sample Report – Predicted Outcomes and Status Changes, February 2022

Member	Total Drug Spend Last 12 Months	Predicted to Be High Drug Spend Member Next Year	Probability of High Drug Cost Member in Next 12 Months	Predicted Total Drug Spend Next 12 Months	New to the Breast Cancer Registry List	Prediction Directionality
1	$60,440	No	0.05	$24,988	No	No Change
2	$53,380	Yes	0.29	$85,773	Yes	New
3	$99,241	Yes	0.56	$118,568	No	Up
4	$78	No	0.01	$5,478	No	No Change
5	$28,454	No	0.28	$56,241	No	Down
6	$169	No	0.14	$38,849	No	Down
7	$301,207	Yes	0.75	$210,807	No	No Change
8	$68,121	Yes	0.61	$145,710	No	Up
9	$10,547	No	0.01	$17,932	No	No Change
10	$28,999	No	0.10	$44,122	No	Up

*Table 4 describes historical drug spending and predicted future drug spending. The new to the breast cancer registry list column indicates Yes if member has never been predicted to be high drug cost and has never had historical drug spend in the last 12 months over $100,000; this should alert care team personnel to members that may not have previously received attention for breast cancer drug costs. Prediction directionality provides a note how the member's prediction is trending from the previous month.

CONCLUSIONS

- Creation of a well-performing breast cancer high drug cost predictive model is possible with integrated medical and pharmacy claims data.
- Breast cancer high drug cost member predictive modeling enhancements with electronic health medical record, laboratory test results, patient reported outcomes, and mortality data should be explored.
- The high drug cost breast cancer member predictive model independent factor identifiers can be used by managed care pharmacists to ensure cost-effective drug therapy optimization.

REFERENCES

1. Blumen H, *et al.* Comparison of Treatment Costs for Breast Cancer, by Tumor Stage and Type of Service. Am Health Drug Benefits. 2016;9(1):23-32. Accessed at www.AHDBonline.com.
2. American Cancer Society. Cancer Facts & Figures 2021. Atlanta: America Cancer Society; 2021. Available from https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2021.html.
3. HighTouchRx. Finding High-Cost Savings for Members with Specialty Conditions. Prime Therapeutics; 2022. Accessed at https://www.primetherapeutics.com/products/hightouchrx/.
4. Symmetry Episode Treatment Groups (ETG) [white paper]. Eden Prairie, MN: OptumInsight, 2022. Accessed at https://www.optum.com/business/insights/health-care-delivery/page.hub.symmetry-etg.html.