

Medicare Member Drug Cost Predictive Model: Creation and Feature Engineering



K. Thompson¹; B.Y. Urick, PharmD, PhD¹; P.P. Gleason, PharmD^{1,2}. ¹Prime Therapeutics LLC, Eagan, MN, United States. ²University of Minnesota College of Pharmacy, Minneapolis, MN, United States.

Background

- Beginning in 2025, out-of-pocket expenses for Medicare Part D beneficiaries will be capped at \$2,000. A total of 11.3 million beneficiaries are projected to meet this cap, and annual savings for these members are expected to average \$600.¹
- This reduction in member cost shares is expected to increase Part D plan liability, potentially increasing premiums and creating incentives for plan sponsors to keep costs below the new spending cap.
- We anticipated that plans would like to understand this potential financial risk and identify potential drug therapy cost-savings opportunities. A predictive model could be used to identify members who are likely to meet the new \$2,000 out-of-pocket threshold. In addition, the predictive model can identify members at risk of having a large increase in their annual drug cost above \$2,000.²

Objective

Our objective was to create a predictive model to identify members enrolled in Medicare Part D who are at risk of large increases in prescription drug spending over the next 12 months, referred to as Rapid Risers, that could help plans manage future pharmacy expenditures.



4085-E 04/25
© 2025 Prime Therapeutics LLC,
2900 Ames Crossing Road,
Eagan, MN 55121

Academy of Managed Care
Pharmacy (AMCP) Annual Meeting,
March 31-April 3, 2025, Houston, TX
Patrick Gleason:
PGleason@PrimeTherapeutics.com
All brand names are property of
their respective owners.

Methods

- The model had two goals: 1) predict members total Part D drug costs in the next 12 months; and 2) identify members whose pharmacy spending is predicted to double over the next 12 months and exceed \$2,000 (referred to as Rapid Risers).
- Two separate models were developed: one for members enrolled in Medicare Advantage prescription drug (MAPD) plans which used integrated pharmacy and medical claims, and one for members enrolled in Medicare Prescription Drug Plans (PDP) which used pharmacy claims only.
- The models were trained on Prime Therapeutics' integrated pharmacy and medical claims data from October 1, 2021, to September 30, 2023, and included 2.5 million Medicare members in the training sample. A range of initial models were tested, including XGBoost, Random Forest, Light GBM, linear regression, and others. Models were developed using machine learning (ML) in the Google Cloud Platform (GCP) using notebooks, pipelines, and other features in the platform,³ and open-source python packages such as sklearn, pandas, and NumPy.
- Model features included in testing included pharmacy and medical claims features such as claim counts and drug costs, eligibility information (including dual eligibility, low-income subsidy, disability, and end-stage renal disease), demographic data, adherence to Medicare Star Rating measures, chronic condition indicators from the Medicare Chronic Condition Warehouse (CCW)⁴, the Charlson Comorbidity Index⁵, the PDP risk score, and diabetes-specific indicators of glucagon-like peptide-1 (GLP-1) use and obesity.
- As high spending is a relatively rare event, techniques included using synthetic minority oversampling technique (SMOTE) to account for the imbalance of the member cost distribution.
- The training/test split of data was set to 85%/15% for data used by the model process, and the latest month of unseen data is used as a holdout for evaluating performance.
- For the model primarily using pharmacy claims, the Random Forest model performed best in the training data for the total Part D drug cost results as judged by root mean squared error (RMSE, lower is better) compared to a model benchmark that uses the member's previous 12 months' drug spending to predict future drug spending.
- For the model using integrated claims, a voting weighted ensemble model consisting of a Random Forest model, a Light GBM model, and an XGBoost model performed best.
- The model output was transformed into a classification model (designation of a Rapid Riser, yes or no). Model performance for the classification results was judged by F5 classification (which prioritized reducing false negative predictions with values from 0 to 1, with closer to 1 being better) compared to a model benchmark that uses the average of all members' previous 12 months' drug spending to predict future drug spending. Using the individual member's previous drug spending is not feasible with this metric since every member would be predicted to not increase in spend and have no predictions of a Rapid Riser.
- Pharmacy claims are refreshed daily, and medical claims are refreshed monthly, so the models often run on different dates of the month to maximize the freshness of the model input data.
- Production results include a monthly report of over 1 million members that includes predictions of Rapid Risers and Part D drug spend in the next 12 months plus other potentially relevant data to end users.

Table 1

Predictive Model Performance Metrics with Comparisons to Benchmark Models

Model	Output	Measure	Range	Predictive Model	Benchmark: Past 12 Months Historical Spending of Person	Benchmark: Past 12 Months Historical Spending of Average Person	Does the Predictive Model Beat the Benchmark?
Pharmacy only	Person's total Part D drug costs in the next 12 months	RMSE*	\$0 to \$infinity – Lower is better	\$10,963	\$11,743	N/A	Yes
Pharmacy only	Rapid Riser ¹ , yes or no	F5-Measure**	0 to 1 – Higher is better	0.488	N/A	0.447	Yes
Integrated	Person's total Part D drug costs in the next 12 months	RMSE*	\$0 to \$infinity – Lower is better	\$10,531	\$11,460	N/A	Yes
Integrated	Rapid Riser ¹ , yes or no	F5-Measure**	0 to 1 – Higher is better	0.455	N/A	0.455	Match

*RMSE: measures the average distance between the predicted values from the model and the actual values in the dataset with an adjustment to penalize larger errors – more sensitive to outliers.

**F5-Measure: Classification metric which favors false positives over false negatives.

¹Rapid Riser is someone whose past 12-month summed pharmacy claims total spend is predicted to double for the future 12 months, and that future 12-month predicted pharmacy spend is over \$2,000.

Table 2

Predictive Model Feature Importance – Top 6 Contributors – Pharmacy Claims Only

Feature Name	Permutation Importance Score
Count of prescribers from pharmacy claims in the past 3 months	0.1220
Total Part D drug spend in the past 2 months	0.0757
Percent of pharmacy claims for multisource generic products in the past 3 months	0.0429
Projected total Part D drug spend in the next 4 months assuming continued full adherence to all drugs	0.0242
Total Part D drug spend in the past 6 months	0.0241
Total Part D drug spend in the past 12 months	0.0171

References

- Assistant Secretary for Public Affairs (ASPA). Inflation Reduction Act and Medicare. US Department of Health and Human Services. Published January 17, 2025. Accessed February 19, 2025. <https://www.hhs.gov/inflation-reduction-act/index.html>
- Cubanski J, Neuman T, Freed M. Explaining the prescription drug provisions in the Inflation Reduction Act. KFF. Published January 24, 2023. Accessed February 19, 2025. <https://www.kff.org/medicare/issue-brief/explaining-the-prescription-drug-provisions-in-the-inflation-reduction-act/>
- Vertex AI platform. Google Cloud. Accessed February 14, 2025. <https://cloud.google.com/vertex-ai?hl=en>
- Chronic Conditions Data Warehouse. Centers for Medicare and Medicaid Services Accessed February 14, 2025. <https://www2.ccmwdata.org/web/guest/home/>
- Glasheen WP, Cordier T, Gumpina R, Haugh G, Davis J, Renda A. Charlson Comorbidity Index: ICD-9 update and ICD-10 translation. Am Health Drug Benefits. 2019;12(4):188-197. <https://pmc.ncbi.nlm.nih.gov/articles/PMC6684052/pdf/ahdb-12-188.pdf>

Table 4

Sample Report – Predictions

Plan Type	Medicare Contract ID	Plan Benefit Package Number	Member ID	Age	Gender	Rapid Riser Indicator	New to Registry Indicator	New Rapid Riser Indicator	Predicted Part D Total Spend Future 12 Months	Part D Total Spend YTD	Part D Total Spend Past 12 Months
MAPD	H0000	123	1	88	F	0	0	0	\$200	\$0	\$0
MAPD	H0000	456	2	92	F	1	0	0	\$2,800	\$0	\$150
MAPD	H0000	123	3	77	M	0	0	0	\$450	\$111	\$167
MAPD	H0000	456	4	91	M	0	1	0	\$395	\$0	\$125
MAPD	H0000	123	5	85	F	1	1	0	\$4,410	\$209	\$458
MAPD	H0000	456	6	68	F	0	0	0	\$155	\$220	\$220
MAPD	H0000	123	7	66	M	1	1	1	\$3,333	\$290	\$451
MAPD	H0000	456	8	77	F	0	0	0	\$355	\$5	\$205
MAPD	H9999	123	9	76	M	0	1	0	\$148	\$235	\$235
MAPD	H9999	456	10	92	F	0	0	0	\$165	\$0	\$97

MAPD=Medicare Advantage Part D; ID=identification number; YTD=year to date

Table 5

Sample Report – Medicare and Clinical Indicators

Member ID	Medicaid Dual Eligible Indicator	Medicare Disability Enrollment Indicator	Low Income Subsidy Indicator	Chronic Condition Warehouse Conditions Past 12 Months ⁴	Proportion Days Medicare Eligible Past 12 Months	Rx Claim Count Past 12 Months	Hospice Indicator Past 12 Months	ESRD Indicator Past 12 Months	Diabetes Indicator Past 12 Months	GLP-1 Indicator Past 12 Months
1	0	0	1	Hyperlipidemia; hypertension; hypothyroidism	1	0	0	0	0	0
2	0	0	0	Rheumatoid arthritis; osteoarthritis; hyperlipidemia; osteoporosis with or without pathological fracture; chronic kidney disease	1	32	0	1	0	0
3	1	1	1	Depression bipolar or other depressive mood disorders	1	10	0	1	0	0
4	0	0	0	Alzheimer's disease; chronic obstructive pulmonary disease	1	7	0	0	0	0
5	0	0	0	Cataract; hyperlipidemia; obesity	1	9	0	0	0	1
6	0	0	0	Stroke transient ischemic attack; hypertension; chronic obstructive pulmonary disease	1	16	0	0	0	0
7	0	1	0	Hyperlipidemia; atrial fibrillation and flutter; obesity	0.5	29	0	0	0	0
8	0	0	0	None	1	6	0	0	0	0
9	0	1	1	Hyperlipidemia	0.25	14	0	0	1	0
10	0	0	0	Hyperlipidemia; hypertension; chronic kidney disease	1	2	0	0	0	0

ID=identification number; ESRD=end-stage renal disease; GLP-1=glucagon-like peptide-1

Results

- For the Random Forest model primarily using pharmacy claims, 199,131 members were evaluated in the test data indexed on October 1, 2023. The test data showed a mean previous total Part D drug spend in the prior 12 months of \$3,210, and median of \$600. Predicted mean Part D drug spend in the next 12 months showed a mean of \$3,739, and a median of \$535.
 - The model outperformed the benchmark for the primary performance metrics, with RMSE of \$10,963 vs. \$11,743, and F5 of 0.488 vs. 0.447.
 - Approximately 26% of the test data were classified as Rapid Risers.
 - Key prediction features identified included previous months' total drug spend, predicted drug cost assuming full adherence, number of prescribers, and multisource generic claim counts.
- For the ensemble model primarily using integrated claims, 126,610 members were evaluated in the test data indexed on October 1, 2023. The test data showed a mean previous total Part D drug spend in the prior 12 months of \$2,976, and median of \$504. Predicted mean Part D drug spend in the next 12 months showed a mean of \$3,684, and a median of \$516.
 - The model outperformed or matched the benchmark for the primary performance metrics, with RMSE of \$10,531 vs. \$11,460, and F5 of 0.456 vs. 0.456.
 - Approximately 25% of members in the test data were classified as Rapid Risers.
 - Key prediction features identified included brand drug utilization, previous months' total drug spend, and specialty claim counts.

Limitations

- Whereas pharmacy claims are updated in the data warehouse daily with almost no lag in time of receipt, medical claims are updated monthly and often lag 1 to 3 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.
- Future improvements could include outputting a predicted list of drugs with drug costs breakouts and predicting when a member could meet their \$2,000 out-of-pocket maximum.
- The model does not incorporate drug rebate discounts.

Conclusions

- A Part D drug cost member predictive model was successfully created using a pharmacy claims only model as well as integrated medical and pharmacy claims data.
- The models performed better than using the members' prior-year drug spend alone.
- The model input used hundreds of features. Features summing drug spending over various time periods contributed prominently in both models. Additionally, the pharmacy claims-only model was influenced by the number of prescribers, and the integrated model was influenced by features related to branded drug spend. We plan to continue investigating new features in future development.
- The model's reporting output, including the drug spend Rapid Riser indicator, medical conditions, and extensive member-specific characteristics, may be used by managed care pharmacists for care management strategies to predict high-cost members and optimize drug therapy in the context of Inflation Reduction Act (IRA)-related Medicare changes.