

Evaluating the Impact of Covariate Lookback Times on Performance of Patient-Level Prediction Models

Jill Hardin (✉ jhardi10@its.jnj.com)

Janssen Research and Development

Jenna M. Reps

Janssen Research and Development

Research Article

Keywords: patient-level prediction, feature extraction lookback periods

Posted Date: February 23rd, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-227607/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

Background: The goal of our study is to provide guidance for deciding which length of lookback to implement when engineering features to use when developing predictive models using observational healthcare data. Using a longer lookback for feature engineering gives more insight about patients but increases the issue of left-censoring.

Methods: We used five US observational databases to develop patient-level prediction models. A target cohort of subjects with hypertensive drug exposures and outcome cohorts of subjects with acute (stroke and gastrointestinal bleeding) and chronic outcomes (diabetes and chronic kidney disease) were developed. Candidate predictors that exist on or prior to the target index date were derived within the following lookback periods: 14, 30, 90, 180, 365, 730, and all days prior to index were evaluated. We predicted the risk of outcomes occurring 1 day until 365 days after index. Ten lasso logistic models for each lookback period were generated to create a distribution of area under the curve (AUC) metrics to evaluate the discriminative performance of the models. Impact on external validation performance was investigated across five databases.

Results: Our results show that a shorter lookback time for the acute outcomes, stroke and gastrointestinal bleeding results in equivalent performance as using longer lookback times. A lookback time of at least 365 days for the chronic outcomes, diabetes and renal impairment, results in equivalent performance as using lookback times greater than 365 days.

Conclusions: Our results suggest the optimal model performance and choice of length of lookback is dependent on the outcome type (acute or chronic). Our study illustrates that use of at least 365 days results in equivalent performance as using longer lookback periods. Researchers should evaluate lookback in the context of the prediction question to determine optimal model performance.

Background

The Observational Health Data Science and Informatics (OHDSI) collaboration have developed an end-to-end framework for developing patient-level prediction (PLP) models [1]. The framework requires data in a standardized data format, the OMOP common data model (CDM) [2], and enables transparent, reproducible and rapid development and validation of prediction models across diverse sets of data allowing for evaluation of previously intractable patient level prediction questions. Briefly the OMOP CDM unifies data from heterogeneous electronic health record and medical insurance claims sources with respect to terminologies and overall structure, allowing us to incorporate data from multiple health care systems into our analysis. The PLP framework applies best practices for model development and evaluation, but there are still subjective choices that need to be made during the model development process. An example is the choice of feature engineering that converts the observational data to the labelled data required for binary classification.

The observational healthcare data consists of timestamped data, which needs to be converted into features for a prediction model. Due to the temporality of the observational data, it is possible to either fully preserve the temporal nature of the data ('temporal features', for example as a feature matrix per patient with rows corresponding to medical events and columns corresponding to time and the entries being the medical event value at the specific time) or create a summary of the patient's history ('non-temporal features', a feature vector per patient corresponding to medical events and the entries are the values, for example binary values indicating the presence or absence of an event in the patient's history). Temporal features can be used with classifiers such as neural networks (deep learning) however, this is not possible with many conventional classifiers (such as logistic regression). In addition, there are difficulties when developing models using temporal data from healthcare claims and electronic healthcare record databases as the data come from a diversity of sources and are recorded at irregular frequencies with data often sparsely represented. This can present issues to classifiers such as neural networks when implementing the feature engineering [3], especially if the data are not large. In this paper we therefore focus on engineering non-temporal features.

Converting observational data to non-temporal data requires specifying a static lookback time where the value of the medical event is observed during the lookback period. It is possible to specify the lookback time, such as 365 days prior to index which means only the data recorded in the 365-days prior to index per patient are used when constructing the features. Alternatively, the lookback window can be specified to include all time prior, meaning all data recorded prior to index are used to construct the features. The benefits of using a longer look back are that you have a more complete picture of each patient, but there are multiple negative aspects including: i) you treat a recent illness the same as an illness experienced years ago, ii) you may have issues with left censoring as patients often do not have the same length of complete lookback and iii) you may run into issues when implementing the model in a new healthcare system if the mean complete lookback is shorter. Figure 1 represents a subject with left censoring (subject A) and a subject without left censoring (subject B). For subject B there is no missing data in the feature construction, but for subject A the left censoring means we are unable to observe her for part of the lookback time (effectively missing data).

[3]Studies using administrative data and investigating variations in the length of lookback period have been conducted in the context of incidence and effect estimation[4–6]. In a study of cancer cumulative incidence estimation the authors recommended using lookback of two or more years and discouraged the use of one year lookback but caveated that it is not possible to provide general recommendations as lookback period is dependent on the characteristics of the cancer site and the available data and the underlying research question[5]. A Korean study using a cohort database and examining lookback and estimating incidence of three gynecological diseases (uterine leiomyoma, endometriosis, and adenomyosis) found that as the lookback increased the proportion of misclassified incident cases decreased but advised that the optimal lookback for annual incidence depended on the nature and the stage of the respective diseases[6]. A comparative effect study using the Medicare beneficiary database and evaluating the effect of statin initiation on incidence of cancer recommended that a three year lookback was best but if infeasible that all available lookback is preferable to short fixed lookbacks[4].

Although these studies do not utilize the PLP methodology they illustrated that longer lookback reduces data noise for the diseases examined.

Few studies have evaluated the impact of the selection of the length of lookback time in the setting of predictive ability [7–9]. In a Korean study with data from the National Health Insurance Database evaluating in hospital mortality for patients aged 40 and older who underwent percutaneous coronary intervention the authors' compared comorbidity measurements (Charlson comorbidity index, Elixhauser's comorbidity, and comorbidity selection) using three years of inpatient records compared to models using one year of inpatient records and concluded the longer lookback period offered no improvement in predictive capacity [8]. Evaluation of the impact of one year vs. two year lookback in Charlson score for mortality among elderly Medicare beneficiaries using claims data reported nearly identical C-statistics [9]. An Australian study using population based hospital data examined prediction of hemorrhage in pregnancy among eight different chronic disease cohorts and evaluated six lookback periods and concluded that although longer ascertainment periods resulted in improvement of identification of chronic disease history it did not change the resulting C-statistics [7]. These studies evaluated a limited set of outcomes (mortality and hemorrhage during pregnancy). Based on the findings of these studies for the outcomes evaluated lookback period did not materially impact the results.

Thus, a systematic evaluation has not been conducted to determine the optimal lookback period for prediction models in the acute and chronic disease areas. The intent of this study is to evaluate the performance of prediction models using acute and chronic disease cohorts and using several lookback periods and multiple databases to provide a recommendation for the optimal lookback period. We hypothesize that using a 365 days prior lookback will result in well performing prediction models that are more transportable across databases as this is a trade-off between gaining a sufficient picture of each patient's health history while reducing issues with left censoring.

Methods

We used the OHDSI PatientLevelPrediction framework[1] and R package to develop and evaluate the prediction models in this study.

Data

We developed models using five US observational datasets. Each dataset has unique attributes. The IBM MarketScan® Commercial Claims and Encounters Database (CCAE) which contains insurance claims for commercially employed individuals and their dependents and contains subjects less than or equal to the age of 65. The Optum® De-Identified Clinformatics® Data Mart Database – Socio-Economic Status – (Optum) is a similar database to CCAE except that it also contains claims from subjects with Medicare supplemental insurance and thus does not have an upper age threshold. IBM MarketScan® Medicare Supplemental and Coordination of Benefits Database (MDCR) database contains claims from subjects with Medicare supplemental insurance and thus contains subjects 65 years and older. The IBM MarketScan® Multi-State Medicaid Database (MDCD) contains claims from subjects covered by

Medicaid and is primarily composed of women and children. The Optum® de-identified Electronic Health Record Dataset (Panther) dataset is an electronic health records (EHR) and contains information derived from clinical Notes using Natural Language Processing (NLP). All databases were transformed to the Observational Medical Outcomes Partnership Common Data Model version 5.3.1. The use of IBM and Optum databases were reviewed by the New England Institutional Review Board (IRB) and were determined to be exempt from broad IRB approval.

For each of the databases ten lasso logistic models for each lookback period (14 days, 30 day, 90 days, 180 days, 365 days, 730 days, and all days prior to index) were generated to create a distribution of area under the curve (AUC) metrics to evaluate the discriminative performance of the models. External validation was performed using one model and validating across four databases.

Study Population

We extracted data for patients who are newly treated with a hypertensive medication to predict four outcomes occurring from 1 day to 365 days after their first prescribed hypertensive treatment.

The target population was new users of hypertensive medications and the eligibility was defined as first time exposure to one or more hypertensive medications on or after 2013 with at least one diagnosis of hypertensive disorder in the 365 days prior to the index drug exposure. We excluded subjects with a prior diagnosis of any of the outcomes evaluated. We required subjects to have at least 365 days of continuous observation prior to the index date. See Appendix A for the codes and logic used to define the target population.

Two of the outcomes were acute health conditions (stroke and gastrointestinal bleeding) and two were chronic health conditions (diabetes and renal impairment) from 1 day after index until 365 days after index. Eligibility for the outcome populations included the first occurrence of gastrointestinal bleeding or stroke or diabetes or renal impairment. See Appendix A for the codes and logic used to define the outcome populations.

Candidate Predictors

Candidate non-temporal features were engineered from the administrative claims data that exist on or prior to the target index date and followed a standardized feature construction process[1]. These variables were demographics, visit type, binary indicators of medical events and counts of record types. The demographics included gender, race and ethnicity (where available), age in 5 year groups (0–4, 5–9, 10–14,..., 95+) and month at the target index date. Binary indicator variables were created for medical events based on the presence or absence of each within the clinical domains of conditions, drugs, and procedures within several time periods: 14, 30, 180, 365, 730 days, and all time prior to index. As binary covariates are the presence or absence of records for various conditions or drugs during time intervals, missing values will not be contained in the covariates. This is because when a patient does not have a condition recorded, we cannot distinguish whether it is due to the patient having the condition but not having it recorded (missing) or them not experiencing the condition. Therefore, missing records for condition, drugs or procedures are treated as the patient not having the condition, drug, or procedure. Left

censoring may result in missing records. Age and gender are mandatory in the OMOP CDM, so are never missing. If a database contains race/ethnicity it will be recorded for all patients.

The published best practices for model development were followed [1]. To enable full transparency and implementation by other researchers using different data all the definitions, analysis code, and prediction models are available in the OHDSI github repository: <https://github.com/ohdsi-studies/PredictionCovariateLookback>.

Development and validation of prediction model

In this study we focused on developing logistic regression with LASSO regularization (LASSO logistic regression) binary classifiers [10]. LASSO logistic regression is a good classifier to use when there is a large number of covariates. To develop the models, 20% test data was sampled from the population and the remaining 80% of the data was used to learn the model (the optimal regularization hyper-parameter was selected using 10-fold cross validation) (Fig. 2). The test data were used to internally validate the model. To evaluate the model's discrimination the area under the receiver operating characteristic curve (AUC) was used. Model development was repeated per lookback period 10 times (using different test/train splits) in order to derive a confidence interval and a mean AUC for the comparisons. External validation was performed on the complete population in the new database.

Results

This study examined model performance over seven lookback periods (14, 30, 90, 180, 365, 730, and all time prior to index) for two chronic (diabetes and renal impairment) and two acute (stroke and gastrointestinal bleeding) outcomes in subjects newly treated with hypertensive medications across five US databases.

Figure 3 shows the internal and external validation discrimination (mean AUC) across the databases for all four models. The rows correspond to the different outcomes and the columns correspond to the database used to develop the model. The two chronic illnesses, diabetes (top row) and renal impairment (third row), show an increasing AUC over the lookback periods up to 365 days (14, 30, 90, 180 days) and then remained stable or improved marginally in CCAE, MDCC, Optum, Panther, and MDCR (renal impairment only). The mean AUC for diabetes models results from MDCR illustrated a continuous increase. For the acute illnesses, the mean AUC for stroke (fourth row) was relatively stable across the lookbacks and for gastrointestinal bleeding (second row) it increased slightly up until 180 days where it stabilized.

The mean number of predictors in the chronic outcome cohorts (diabetes and renal impairment) increase with lookback time and the largest number of predictors were observed in CCAE and the smallest number of predictors were found in MDCR (Fig. 4). The acute outcome cohorts (stroke and gastrointestinal bleeding) illustrate the same pattern of mean number of predictors to the chronic outcome cohort models however overall fewer numbers of predictors are included in the final models.

Discussion

The objective of this study was to provide empirical evidence to inform selection of non-temporal feature lookback windows by investigating two acute and two chronic disease outcomes.

Review of our study findings

The main findings of our results are:

1. The choice of lookback made little impact in terms of discriminative performance, although the performance did generally appear worse when a lookback of 180 days or less is used.
2. The trend of the impact of lookback time on internal validation appears to match the trend of the impact of lookback on external validation in the databases investigated.
3. A lookback of 365 days may be a good trade-off between maximizing discriminative ability and minimizing the model complexity.

Overview of previous work

Studies of lookback periods have been conducted in the context of incidence estimation/phenotype development, effect estimation, with a limited number of studies in the context of prediction. Incidence estimation and effect estimation studies all advise that longer lookback reduces data noise for the diseases and data sources evaluated. The findings from the relatively few prediction studies stated that for the outcomes evaluated that lookback period did not materially impact the model discrimination but did improve identification of the disease history.

Strengths and Limitations

This is the first study that systematically examined the impact of lookback period for two acute and two chronic outcomes on PLP model discrimination using administrative claims data. The data sources used afforded large sample sizes and long observation periods resulting in good precision and the ability to evaluate several lengths of lookback period.

Our study focuses on two examples of chronic and acute outcomes, utilizes US administrative claims data, and uses a single prediction algorithm, the LASSO logistic regression. There is no guarantee that the trends observed in this study would generalize across all outcomes, models and data. Therefore, future research should focus on evaluation of additional outcomes, utilize alternative types of data sources, and evaluate additional prediction algorithms. We recommend that future PLP studies evaluate more than one lookback period to select the lookback that results in optimal discrimination for a given prediction question.

The algorithms used to identify the four outcomes likely are prone to some misclassification, although we performed no formal evaluation of the operating characteristics. We did inspect cohort definitions and characteristics using the CohortDiagnostics R package [11] prior to execution of the PLP models. Our study evaluated seven lookback periods but none in combination and thus it is possible that

combinations of lookback periods could result in model discrimination improvement. Model performance utilizing different lookbacks may not be generalizable to alternative target and outcome populations, data sources, and prediction algorithms.

Conclusions

The results of our study suggest for the two chronic outcomes evaluated that a lookback of at least 365 days be evaluated and that for the acute outcomes evaluated that a lookback of ≤ 365 days was sufficient to optimize model discrimination. However, the selection of the length of lookback period, the type of outcome evaluated, the characteristics of data source utilized, and the underlying research question impacts model discrimination and therefore it is not possible to provide general recommendations. We advise prediction studies using administrative data consider several choices of lookback to achieve optimal model discrimination and robust results.

List Of Abbreviations

OHDSI = Observational Health Data Science and Informatics

PLP = patient-level prediction

OMOP = Observational Medical Outcomes Partnership

CDM = common data model

CCAIE = Commercial Claims and Encounters Database

MDCR= Medicare Supplemental and Coordination of Benefits Database

MDCD = Multi-State Medicaid Database

EHR = electronic health records

NLP = Natural Language Processing

IRB = Institutional Review Board

AUC = area under the curve

Declarations

Ethics approval and consent to participate

The use of IBM and Optum databases were reviewed by the New England Institutional Review Board (IRB) and were determined to be exempt from broad IRB approval.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from IBM MarketScan® and Optum® but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data are however available from the authors upon reasonable request and with permission of IBM MarketScan® and Optum®. To enable full transparency and implementation by other researchers using different data all the definitions, analysis code, and prediction models are available in the OHDSI github repository: <https://github.com/ohdsi-studies/PredictionCovariateLookback>.

Competing interests (same as provided on the submission system)

Both authors (Jill Hardin, Jenna Reps) are employees of Janssen Research & Development, LLC, and may own stock and/or stock options. The work on this study was part of their employment.

Funding

Funding for this research was provided by Janssen Research & Development, LLC. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Authors' contributions

All authors contributed to all aspects (study design and execution, data analysis and interpretation, and writing of the manuscript) of the study.

Acknowledgements

The research was supported by Janssen Research & Development, LLC.

Authors' information

Not applicable.

References

1. Reps, J.M., Schuemie, M.J., Suchard, M.A., Ryan, P.B. and Rijnbeek, P.R., *Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data.* . J Am Med Inform Assoc. , 2018. **25**(8): p. 969-975.
2. Overhage JM, R.P., Reich CG, Hartzema AG, Stang PE, *Validation of a common data model for active safety surveillance research.* J Am Med Inform Assoc., 2011. **19**(1): p. 54-60.
3. Moskovitch, R., et al., *Temporal biomedical data analytics.* J Biomed Inform, 2019. **90**: p. 103092.

4. Conover, M.M., et al., *Classifying medical histories in US Medicare beneficiaries using fixed vs all-available look-back approaches*. *Pharmacoepidemiol Drug Saf*, 2018. **27**(7): p. 771-780.
5. Czwikla, J., K. Jobski, and T. Schink, *The impact of the lookback period and definition of confirmatory events on the identification of incident cancer cases in administrative data*. *BMC Med Res Methodol*, 2017. **17**(1): p. 122.
6. Kim, M., et al., *The effect of the look-back period for estimating incidence using administrative data*. *BMC Health Serv Res*, 2020. **20**(1): p. 166.
7. Chen, J.S., et al., *Use of hospitalisation history (lookback) to determine prevalence of chronic diseases: impact on modelling of risk factors for haemorrhage in pregnancy*. *BMC Med Res Methodol*, 2011. **11**: p. 68.
8. Kim, K.H. and L.S. Ahn, *[A comparative study on comorbidity measurements with Lookback period using health insurance database: focused on patients who underwent percutaneous coronary intervention]*. *J Prev Med Public Health*, 2009. **42**(4): p. 267-73.
9. Zhang, J.X., T.J. Iwashyna, and N.A. Christakis, *The performance of different lookback periods and sources of information for Charlson comorbidity adjustment in Medicare claims*. *Med Care*, 1999. **37**(11): p. 1128-39.
10. Suchard, M.A., et al., *Massive parallelization of serial inference algorithms for a complex generalized linear model*. *ACM Trans Model Comput Simul*, 2013. **23**(1).
11. Rao, G., Schuemie, M. , Ryan, P. , Weaver, J. *CohortDiagnostics*. 2020; Available from: <https://ohdsi.github.io/CohortDiagnostics/>.

Figures

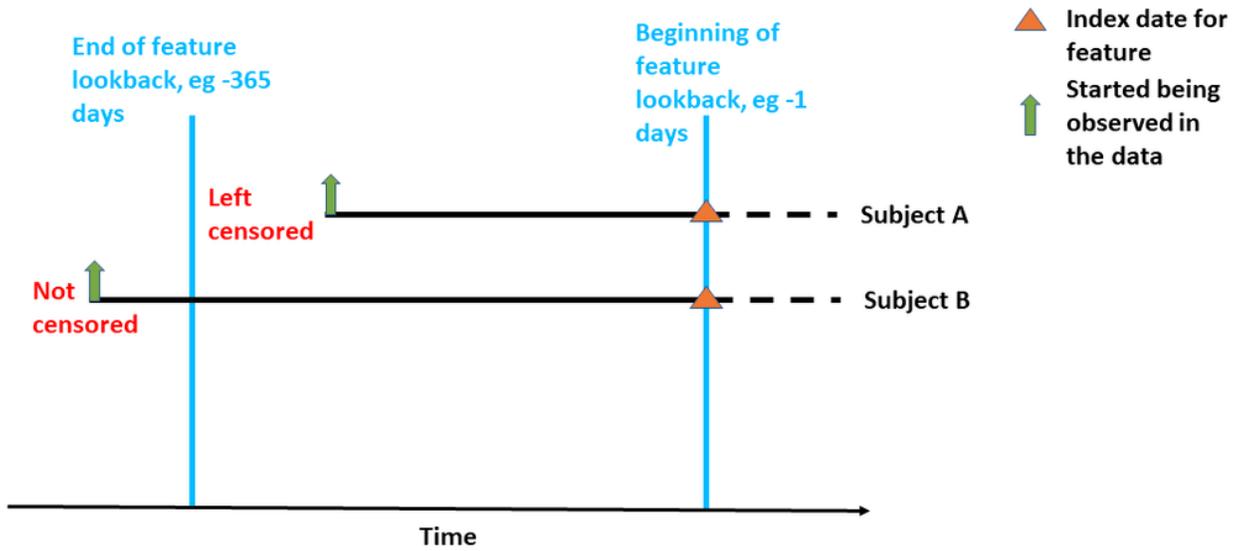


Figure 1

Left censoring and feature construction

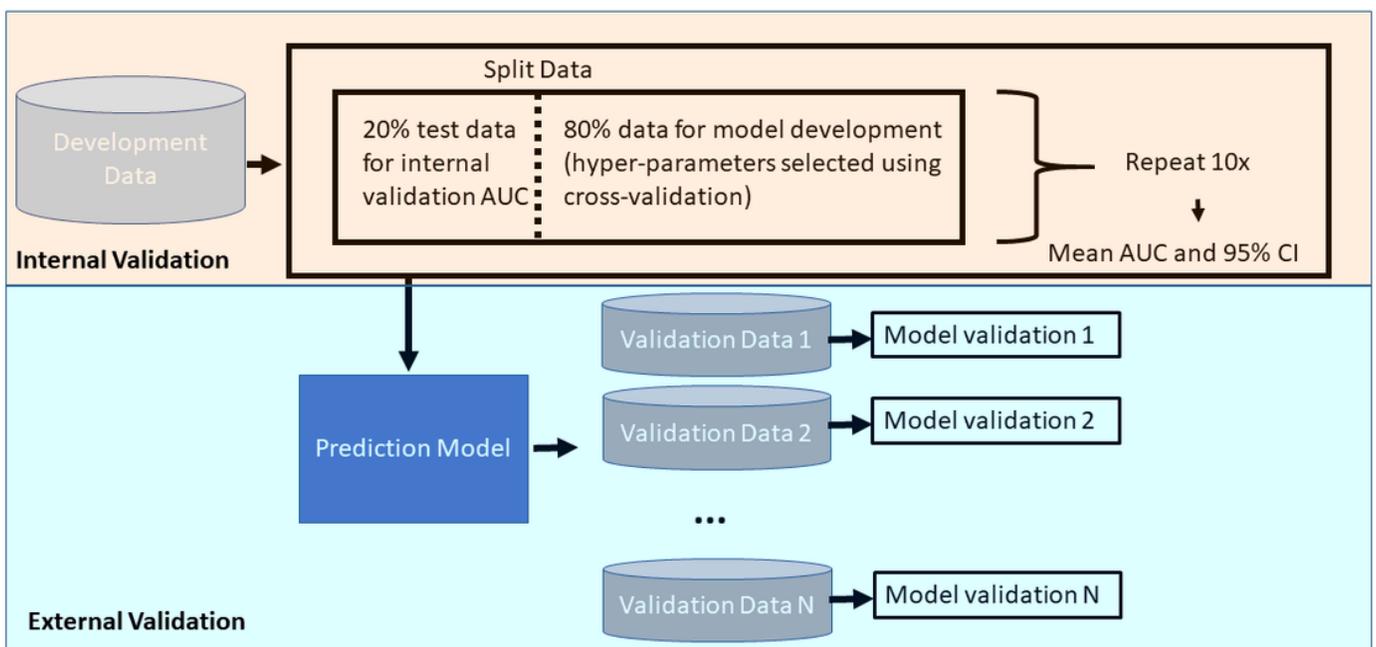


Figure 2

36. Model development and internal and external validation process

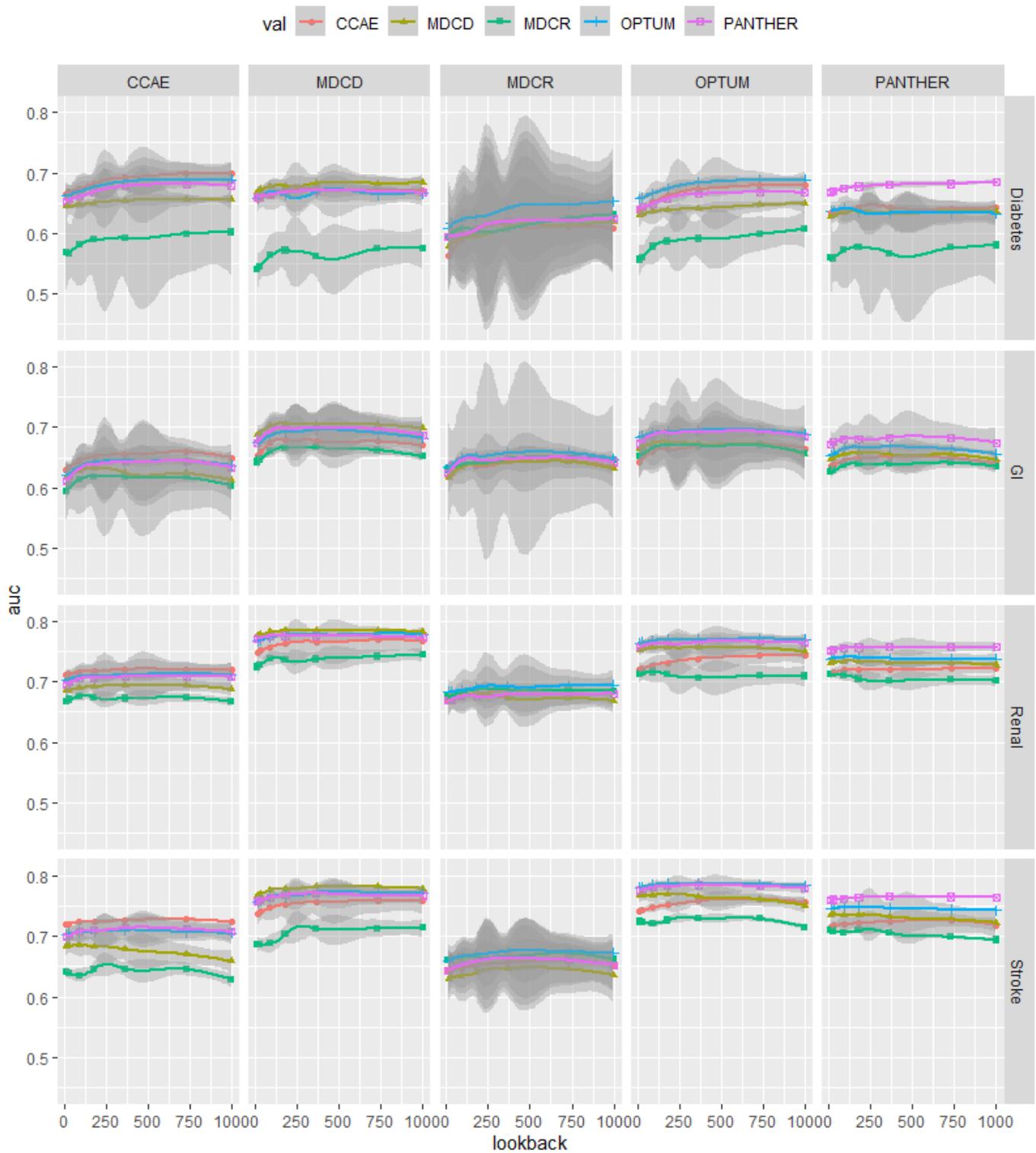


Figure 3

Mean AUC across two chronic (Diabetes and Renal Impairment) and two acute (Gastrointestinal bleeding and Stroke) cohorts over five US databases. The database at the top was used to train the model. Colors

and shapes indicate the database used to compute the AUC. The shaded area indicates the range of AUCs observed across the 10 replications.

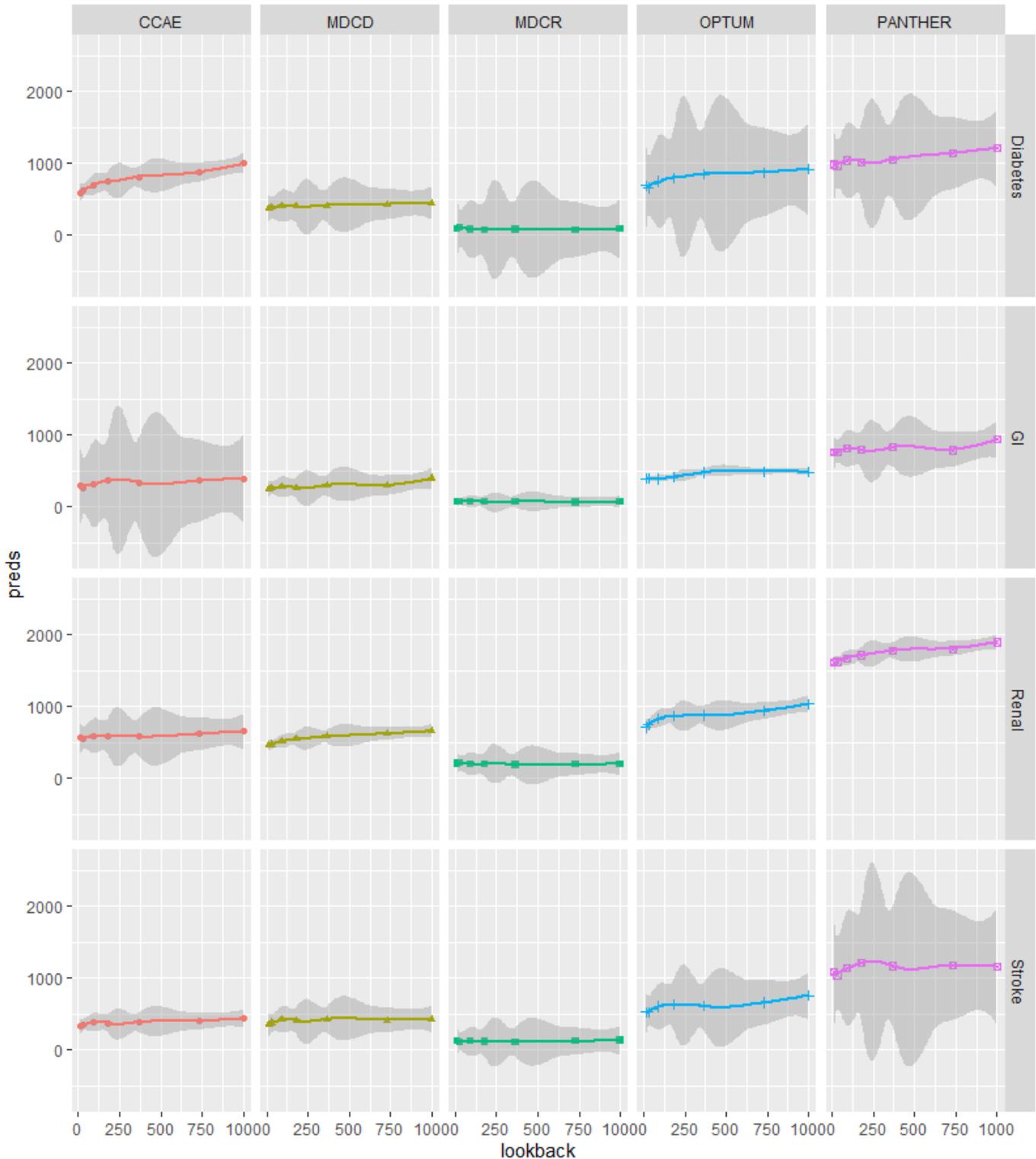


Figure 4

Mean predictors across two chronic (Diabetes and Renal Impairment) and two acute (Gastrointestinal bleeding and Stroke) cohorts over five US databases.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [HardinAppendixA.xlsx](#)
- [HardinAppendixB.xlsx](#)
- [HardinSupplementaryFileTitleAuthorList.docx](#)