Best practices for prognostic model development using observational health data: a scoping review

Cynthia Yang[1], Jenna M. Reps[2]
[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Observational Health Data Analytics, Janssen Research and Development, Titusville, NJ, USA

Background

Personalizing medical decision making by incorporating patient-level risk estimates could improve patient health. Prediction models also have the benefit of removing subjectivity by enabling clinicians to apply a tool that provides standardized and objective risk scores. Prediction models can be developed using different types of medical data including medical images, laboratory data, clinical trial data, survey data and observational health data such as insurance claims and electronic health record (EHR) data. Observational health data often contain more patients, and these patients are often more representative of the general population. However, a disadvantage of observational data is that they are generally not being collected primarily for the purpose of research.

In our previously published paper, we demonstrated how the OHDSI PatientLevelPrediction framework enables a standardized analytics pipeline for the development and validation of clinical prediction models across observational databases (1, 2). Developing a prediction model requires careful consideration of the prediction problem of interest and study design choices such as sensitivity analysis options. Initial study design choices should be guided by best practices to improve research quality (3). While several papers propose best practices for prediction model development (4-7), it is important to ensure these best practices are also suitable for big observational health data. The aim of this scoping review is to present an overview of current knowledge and empirical evidence supporting best practices for prognostic model development using observational health data.

Methods

We conducted a preliminary PubMed search for relevant papers published in the period 2018-2022. We included all papers that empirically investigated the impact of study design choices on the performance and interpretation of prognostic prediction models using real data. We used the following search query on June 13, 2022:

(((patient-level prediction) OR (patientlevelprediction) OR (patient level prediction) OR (prognostic))
AND ((empirically) OR (empirical) OR (investigate) OR (impact))
AND (claims OR ehr or records or observational))
OR
(prediction AND (empirically investigate) AND design).

JMR screened all titles to identify potentially relevant papers, after which CY and JMR assessed eligibility of all potentially relevant papers based on the full text. Data extraction was completed by CY. We extracted the following information: research topic, research aim, whether OMOP CDM was used, type of data, number of prediction tasks, classifier, and recommendations. We categorized the research topics based on the following steps in the process required to develop a prediction model for a specific task from observational health data:

1) Data Extraction: how to extract the longitudinal observational health data into tabular form
2) Data Pre-processing: how to pre-process the tabular data
3) Model development: which classifiers and hyper-parameter search strategy
4) Model validation: how to assess the model's performance (which metrics, internal vs external validation)

Preliminary results

The PubMed search resulted in a total of 6,671 papers. From this, 36 potentially relevant papers were identified. Upon full text inspection, 11 papers were eventually included for synthesis (Table 1). We identified 2/11 papers on

data extraction, 3/11 papers on data pre-processing, 3/11 papers on model development, and 1/11 papers on model validation. 2/11 papers covered multiple research topics. 6/11 papers used data mapped to the OMOP CDM, and 3/5 remaining papers used data from MIMIC-III (among others). All papers included some form of logistic regression as one of the classifiers, while 6/11 papers also used other classifiers for their empirical investigations.

Table 1. Overview of research on best practices in the period 2018-2022

Research topic	Research aim	OMOP CDM used?	Type of data	Number of prediction tasks	Classifier	Recommendations
Data extraction	To examine the impact of the lookback period (prior to index) when creating covariates (8).	Yes	Four US claims and one US EHR databases	Four different outcomes: acute (stroke and gastrointestinal bleeding) and chronic outcomes (diabetes and chronic kidney disease) in patients with hypertensive drug exposures	Lasso logistic regression	A short lookback (< 180 days) can limit a model's performance (in terms of AUROC), but there was only a small gain in performance going back further than one year. A one-year lookback period seems to be a good trade-off between performance and interpretability.
Data extraction	To systematically assess the impact of diagnosis code groupings as covariates: AHRQ-Elixhauser, Single-level CCS, truncated ICD-9-CM codes, and raw ICD-9-CM codes (9).	No	Two US EHR databases: 1) MIMIC-III, 2) University Health System	Three prediction tasks: 1-year mortality following an ICU stay, 30-day mortality following surgery, and 30-day complication following surgery	Lasso logistic regression, ridge regression, random forest, and gradient boosting	Single-level CCS groupings represent aggregations of raw codes into meaningful clinical concepts and consistently balance interoperability between ICD-9-CM and ICD-10-CM while maintaining strong model performance as measured by AUROC and AUPRC.
Data pre-processing	To investigate the impact of the decision of how to include or exclude patients who are lost to follow-up (10).	Yes	One US claims database	21 different outcomes in depression	Lasso logistic regression and lasso cox regression	Excluding patients without the outcome who are lost to follow-up while including patients with the outcome who are then lost to follow-up leads to model bias. Either include Either include or exclude all patients lost to follow-up.
Data pre-processing	To determine the sample size at which near optimal performance can be achieved (11).	Yes	Three US claims databases	23 outcomes in depression, 58 outcomes in hypertension	Lasso logistic regression	In most cases only a fraction of the available data was sufficient to produce a model close to the performance of one developed on the full data set, while substantially reducing model complexity.
Data pre-processing	To explore the impact of imputation methods on model performance and the derived interpretations (12).	No	One US EHR database: MIMIC-III	One prediction task: all-cause mortality among acute myocardial infarction patients	Logistic regression, support vector machine, and decision tree	GAIN and MissForest yielded the best imputation performance (in terms of RMSE and small standard deviations) across five levels of missingness. However, variance in subsequent prediction performance (in terms of AUROC) gradually grows with more missingness, and similarity of feature importance of models based on the imputed data to the feature importance of baseline models gradually decreases.
Model development	To determine whether ensembles that combine models developed independently using different databases can improve model transportability (13).	Yes	Four US claims and one US EHR databases	21 different outcomes in depression	Lasso logistic regression	A simple federated learning approach that implements ensemble techniques to combine models developed independently across different databases for the same prediction question may improve the discriminative performance in new data but will need to be recalibrated using the new data.
Model development	To propose a multicenter collaborative prediction model construction framework to build a model with greater generalizability and continuous improvement capabilities while preserving patient data security and privacy (14).	Yes	Five US datasets and one Chinese hospital-specific colorectal cancer dataset	One prediction task: colorectal cancer prognosis	Logistic regression	A multicenter collaborative prediction model construction framework can support the construction of prediction models with better generalizability and continuous improvement capabilities without the need to aggregate multicenter patient-level data.

Model development	To explore the added value of flexible machine learning (ML) algorithms to traditional regression approaches (15).	No	1) IMPACT-II database, 2) CENTER-TBI core study	One prediction task: 6-month mortality and unfavorable outcome after moderate or severe traumatic brain injury	Logistic regression, lasso logistic regression, ridge regression, support vector machine, neural network, random forest, and gradient boosting	In a low-dimensional setting, flexible ML algorithms do not perform better than more traditional regression models in outcome prediction after moderate or severe traumatic brain injury.
Model validation	To determine the impact of the choice of development and internal validation design on the internal performance bias and model generalizability in big data (16).	Yes	Three US claims databases	21 different outcomes in depression	Lasso logistic regression	To limit overfitting: (1) use a hold-out set to pick any hyperparameters (e.g., a validation set or CV) and (2) use a hold-out set to evaluate the model internally (e.g., a test set or CV).
Data extraction, data pre-processing, model development, model validation	To evaluate the impact of model complexity, sample size, prediction periods and training, and validation strategies (17).	No	Two cohorts: one at the University of Kiel and one at the University of Greifswald	One prediction task: tooth loss prediction in patients with periodontitis	Logistic regression, recursive partitioning, random forest, and gradient boosting	More complex models (random forest, gradient boosting) had no consistent advantages over simpler ones (logistic regression, recursive partitioning). Internal validation overestimated the AUROC, while external validation found lower AUROC. Reducing the sample size decreased the predictive power, particularly for more complex models. Censoring the prediction period had limited impact.
Data pre-processing, model development, model validation	To compare approaches formulated to improve disaggregated and worst-case model performance over subpopulations (through modifications to training objectives, sampling approaches, or model selection criteria) with standard approaches for learning predictive models (18).	No	Three EHR databases: 1) MIMIC-III, 2) eICU, 3) STARR	Three different outcomes: in-hospital mortality, prolonged length of stay, and 30-day readmission	Fully connected feed-forward networks, gated recurrent units, and logistic regression	When it is of interest to improve model performance for patient subpopulations beyond what can be achieved with standard practices, it may be necessary to do so via data collection techniques that increase the effective sample size or reduce the level of noise in the prediction problem.

Abbreviations: area under the receiver operator characteristic curve (AUROC), area under the precision recall curve (AUPRC), root-mean-square-error (RMSE), cross-validation (CV).

Conclusion

This scoping review provides an overview of the empirical research that has been done in the period 2018-2022 on the impact of study design choices on the performance and interpretation of prognostic prediction models. Most of the research was done using databases mapped to the OMOP CDM or using the MIMIC-III database. It could be interesting to extend the empirical investigations done using the MIMIC-III database to databases mapped to the OMOP CDM. It could also be interesting to extend the empirical investigations based on only logistic regression to other classifiers. An advantage of using the OMOP CDM is that empirical investigations can easily be done across multiple databases or classifiers to increase generalizability of the recommendations for best practices. Overall, we believe there are a lot of study design choices for prognostic model development using observational health data that still require more research. Examples of useful future research include the following:

- What is the impact of phenotypes? Defining an algorithm to identify patients within a database who have a certain condition or medical event is known as 'phenotyping'. In general, a phenotype can be defined as an index rule followed by inclusion/exclusion rules. It could be useful to investigate the impact of mis-specifying the target population and/or outcome phenotype definitions.
- What is the impact of over/under-sampling? Many datasets used to develop prediction models are imbalanced. In the machine learning literature, it has been suggested that developing models using resampled data may improve prediction performance. However, a recent study suggests that resampling data may worsen performance of the developed prediction models and should therefore not be applied (19). More research needs to be done to empirically investigate the impact of resampling methods on the development and validation of prediction models using observational health data.
- Should you do deep learning to automatically do feature engineering? Deep learning models may be able

to perform better than other machine learning models if it can learn useful patterns from temporal data. It would be useful to compare standard machine learning models using fixed lookbacks periods for the covariates with deep learning techniques that use temporal features.

- How should we interpret external validation performance? If a model trained in database 1 performs poorly on database 2, does this mean the model is bad? The decrease in performance could be due to the model being overfit to database 1, which is problematic. However, it could also be due to database 2 missing important variables or the case-mix differing. It would be useful to have methods that can put external validation into context and can explain the external validation performance.
- Recalibration methods? Model updating? Many studies have highlighted the issue with models being miscalibrated when they are transported into a new setting (new database or different target population). There is a need to learn best approaches for recalibration (e.g., how much labelled data are needed for recalibration or is there a way to recalibrate with just knowledge of the outcome rate in the new data)?
- Is there a way to automatically simplify models? The more covariates a model has, the more difficult to implement. Therefore, is there an automatic process that can be applied after model development to reduce the model covariates while maintaining adequate performance?

References

1. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. 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):969-75.
2. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernandez-Bertolin S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed. 2021;211:106394.
3. Van Calster B, Wynants L, Riley RD, van Smeden M, Collins GS. Methodology over metrics: current scientific standards are a disservice to patients and society. J Clin Epidemiol. 2021;138:219-26.
4. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) The TRIPOD Statement. Circulation. 2015;131(2):211-9.
5. Hemingway H, Croft P, Perel P, Hayden JA, Abrams K, Timmis A, et al. Prognosis research strategy (PROGRESS) 1: a framework for researching clinical outcomes. BMJ. 2013;346:e5595.
6. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis Research Strategy (PROGRESS) 2: prognostic factor research. PLoS Med. 2013;10(2):e1001380.
7. Steyerberg EW, Moons KG, van der Windt DA, Hayden JA, Perel P, Schroter S, et al. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. PLoS Med. 2013;10(2):e1001381.
8. Hardin J, Reps JM. Evaluating the impact of covariate lookback times on performance of patient-level prediction models. BMC Med Res Methodol. 2021;21(1):180.
9. Kansal A, Gao M, Balu S, Nichols M, Corey K, Kashyap S, et al. Impact of diagnosis code grouping method on clinical prediction model performance: A multi-site retrospective observational study. Int J Med Inform. 2021;151:104466.
10. Reps JM, Rijnbeek P, Cuthbert A, Ryan PB, Pratt N, Schuemie M. An empirical analysis of dealing with patients who are lost to follow-up when developing prognostic models using a cohort design. BMC Medical Informatics and Decision Making. 2021;21(1):43.
11. John LH, Kors JA, Reps JM, Ryan PB, Rijnbeek PR. Logistic regression models for patient-level prediction based on massive observational data: Do we need all data? Int J Med Inform. 2022;163:104762.
12. Payrovnaziri SN, Xing A, Salman S, Liu X, Bian J, He Z. Assessing the Impact of Imputation on the Interpretations of Prediction Models: A Case Study on Mortality Prediction for Patients with Acute Myocardial Infarction. AMIA Jt Summits Transl Sci Proc. 2021;2021:465-74.
13. Reps JM, Williams RD, Schuemie MJ, Ryan PB, Rijnbeek PR. Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. BMC Medical Informatics and Decision Making. 2022;22(1):142.
14. Tian Y, Chen W, Zhou T, Li J, Ding K, Li J. Establishment and evaluation of a multicenter collaborative prediction model construction framework supporting model generalization and continuous improvement: A pilot

study. Int J Med Inform. 2020;141:104173.
15. Gravesteijn BY, Nieboer D, Ercole A, Lingsma HF, Nelson D, van Calster B, et al. Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury. J Clin Epidemiol. 2020;122:95-107.
16. Reps JM, Ryan P, Rijnbeek PR. Investigating the impact of development and internal validation design when training prognostic models using a retrospective cohort in big US observational healthcare data. BMJ Open. 2021;11(12):e050146.
17. Krois J, Graetz C, Holtfreter B, Brinkmann P, Kocher T, Schwendicke F. Evaluating Modeling and Validation Strategies for Tooth Loss. J Dent Res. 2019;98(10):1088-95.
18. Pfohl SR, Zhang H, Xu Y, Foryciarz A, Ghassemi M, Shah NH. A comparison of approaches to improve worst-case predictive model performance over patient subpopulations. Sci Rep. 2022;12(1):3254.
19. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. Journal of the American Medical Informatics Association. 2022.