Reps *et al.*
Diagnostic and Prognostic Research (2025) 9:10
https://doi.org/10.1186/s41512-025-00191-x

Diagnostic and Prognostic Research

RESEARCH

Open Access

Can we develop real-world prognostic models using observational healthcare data? Large-scale experiment to investigate model sensitivity to database and phenotypes

Jenna M. Reps[1,2*], Peter R. Rijnbeek[2] and Patrick B. Ryan[1]

Abstract

Background Large observational healthcare databases are frequently used to develop models to be implemented in real-world clinical practice populations. For example, these databases were used to develop COVID severity models that guided interventions such as who to prioritize vaccinating during the pandemic. However, the clinical setting and observational databases often differ in the types of patients (case mix), and it is a nontrivial process to identify patients with medical conditions (phenotyping) in these databases. In this study, we investigate how sensitive a model's performance is to the choice of development database, population, and outcome phenotype.

Methods We developed > 450 different logistic regression models for nine prediction tasks across seven databases with a range of suitable population and outcome phenotypes. Performance stability within tasks was calculated by applying each model to data created by permuting the database, population, or outcome phenotype. We investigate performance (AUROC, scaled Brier, and calibration-in-the-large) stability and individual risk estimate stability.

Results In general, changing the outcome definitions or population phenotype made little impact on the model validation discrimination. However, validation discrimination was unstable when the database changed. Calibration and Brier performance were unstable when the population, outcome definition, or database changed. This may be problematic if a model developed using observational data is implemented in a real-world setting.

Conclusions These results highlight the importance of validating a model developed using observational data in the clinical setting prior to using it for decision-making. Calibration and Brier score should be evaluated to prevent miscalibrated risk estimates being used to aid clinical decisions.

Introduction

Large observational healthcare databases, such as insurance claims data or electronic healthcare data, can be used to develop health prediction models that get implemented in various real-world settings [1]. For example, QRISK is a model that was trained using data from a subset of primary care practices in the UK. The model predicts 10-year risk of cardiovascular disease and has been implemented in the UK to identify patients who may benefit from risk-lowering interventions such as initiating statins [2]. The Revised Cardiac Risk Index (RCRI) is

*Correspondence:
Jenna M. Reps
jreps@its.jnj.com
[1] Johnson & Johnson, Raritan, NJ, USA
[2] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

a model that is commonly applied preoperatively to estimate the perioperative risk of cardiovascular complications [3]. Recently, when the COVID pandemic started, researchers used these databases to develop models to identify which patients were at highest risk of serve COVID outcomes [4–6]. These COVID models were developed using observational data, but the aim was often to apply the models in clinical settings to identify which patients should have been prioritized a COVID vaccination or other forms of intervention [7].

Targeted validation is a term that represents validating a model in the intended setting [8]. This is an important consideration for models developed using observational healthcare data that will be implemented in real-world clinical settings. Observational healthcare databases may not contain a representative sample of the intended clinical population (i.e., if a model is developed using US insurance claims data that includes patients who are commercially employed, these patients are likely to be younger and healthier than the average US population). In addition, it is not always clear how to identify patients with medical conditions in observational healthcare data (rule-based criteria are often used, but these are subjective, and chart review to assess accuracy is not always feasible in claims data). When a model developed using observational healthcare data is transported into a clinical setting, the model performance can be impacted by numerous factors including (i) change in patient case mix, (ii) missing predictors, (iii) predictors that have different meaning, (iv) change in implementation timing (e.g., applying the model at the first visit of the year vs at the first record of some medical condition), (v) outcome difference (i.e., the model was developed using patients with more or less severe outcomes, or there was differential measurement error in the outcome between the training data and targeted application). Prior studies have shown a deterioration in performance when models are transported across databases [9]. However, there has been a lack of research into the impact that these different factors, and their interactions, have on transported model performance.

In particular, there has been little research into how much the population and outcome definitions impact a model's performance stability. A large amount of work is often required by researchers to develop rule-based criteria or models that aim to identify patients with medical conditions of interest in observation healthcare datasets [10]. This process is known as phenotyping, and the definition is referred to as a phenotype definition. For example, Cai et al. used the rule-based phenotype of having a diagnosis ICD- 9 code: 410.XX in the primary position during an inpatient visit for identifying acute myocardial infarction [11]. However, it is common for different researchers to use different phenotype definitions for the same medical condition, as seen for acute myocardial infarction [12]. It is of interest to determine how sensitive a prediction model is to the choice of population and outcome phenotype. If a model's performance varies substantially based on these choices, then extensive work needs to be performed to ensure accurate phenotypes, which match the intended target population, are used.

In this study, we develop and validate prediction models for nine prediction tasks across seven observational databases, three populations with different prediction indexes, and three to four outcome definitions per task to investigate model stability. We investigate stability by permuting the database, population, and/or outcome definition. Stability is investigated in terms of overall performance metrics (discrimination, calibration, and Brier score) and individual predicted risks [13]. This serves as a proxy for how stable model performance may be between the development database performance and targeted validation performance. It will provide insight into the impact that choice of development database, incorrectly defining populations or using noisy outcomes, has on model stability.

Methods

Aims

We aim to investigate how stable model performance is when there is a change to the following:

1) The population (e.g., the model is developed using a development population consisting of patients with an outpatient visit in 2017, index is first visit, and then the model is validated using a population of patients observed in the database during 2017, but index is Jan 1, 2017). Changing the population impacts case mix and implementation timing.
2) The outcome definition (e.g., the model is developed using rule-based criteria to identify acute myocardial infarction that consists of a patient having a single code representing acute myocardial infarction but is applied using rule-based criteria requiring a patient to have a single code representing acute myocardial infarction during an inpatient visit). Changing the outcome definition impacts the outcome measurement error.
3) The development database (e.g., the model is developed in database 1 and validated in database 2). Changing the database impacts the case mix, the prediction meaning, and which predictors are available.
4) The population, outcome definition, and/or database to investigate the interaction between these factors on model stability.

We will measure model stability using previously defined stability levels: level 1 (comparing change in performance metrics) and level 4 (comparing changes in individual risk estimates for a selection of patients) [13].

Data

We use seven observational databases in this study. All databases are converted into the Observational Medical Outcomes Partnership (OMOP) common data model (CDM) format [14] that use standardized vocabularies for recording medical events. Full database details are available in Table 1.

The use of JMDC, IQVIA, Merative MarketScan®, and Optum databases were reviewed by the New England Institutional Review Board and were determined to be exempt from broad Institutional Review Board approval.

Prediction task

In this study, we follow the proposed prediction framework by Reps et al. [15].

To determine how stable model performance is, we focus on the prediction tasks of developing models that predict various COVID- 19 vaccine outcomes of interest listed by the US Food and Drug Administration (FDA) for the general population (as vaccines are available for every adult in the USA). We chose these tasks as the prediction models need to be valid for the whole adult population, but this makes the choice of development population and database unclear. We chose a population of non-COVID vaccine users since when these models would have been useful, there would have been little data on COVID vaccinations in observational data. Using a proxy target population was common at the beginning of the pandemic [5].

We chose three realistic (researchers may use them) populations for the tasks:

1) Patients with a recorded healthcare visit in 2017 with 365 days or more observation time in the database prior to the visit. Index was the first of these visits per patient. This population is suitable if the model would be applied during a patient's first healthcare interaction in any year.
2) Patients with a recorded influenza vaccine in 2017 with 365 days or more observation time in the database prior to the influenza vaccine. Index was vaccine date. This population is suitable if the model would be applied when a patient is vaccinated each year.
3) Patients observed in the database during 2017 with 365 days or more observation time in the database prior to Jan 1, 2017. Index was Jan 1, 2017. This population is suitable if the model would be applied every year on the 1 st of January.

We restricted to the year 2017 to avoid using data from 2019 onwards as this was impacted by the pandemic. As some of the populations were large, we took a random sample of 2 million patients when the population was greater than 2 million patients.

We focus on nine prediction tasks, specifically predicting the first-time occurrence of nine outcomes from 1 day until 365 days after index. The nine outcomes were as follows: acute myocardial infarction (acute MI), anaphylaxis, appendicitis, disseminated intravascular coagulation (disintracoag), encephalomyelitis, Guillain–Barre syndrome, hemorrhagic stroke, nonhemorrhagic stroke, and pulmonary embolism. These outcomes were chosen as they have been identified as COVID- 19 vaccine outcomes of interest by the FDA, and we wanted to see whether observational healthcare databases could be used to predict these outcomes for the general population. For each prediction task, we developed models using three different outcome phenotype definitions (except nonhemorrhagic stroke that had four). In general, the first phenotype was a standardized definition that looked for at least one occurrence of a diagnosis code corresponding to the clinical idea, as defined by the OHDSI standardized vocabularies. The second phenotype identified outcomes based on occurrence of a

Table 1 The databases used in this study

Database full name	Database short name	Type	Size (million patients)
Merative Commercial Claims and Encounters	CCAE	US insurance claims	173
IQVIA Disease Analyzer — Germany	IQVIA_Germany	German primary care	33
Japan Medical Data Center	JMDC	Japanese insurance claims	19
Merative Medicaid	MDCD	US insurance claims	37
Merative Medicare Supplemental Beneficiaries	MDCR	US insurance claims	11
Optum® De-Identified Clinformatics® Data Mart Database	Optum Clinformatics®	US insurance claims	101
Optum® De-identified Electronic Health Record Dataset	Optum® EHR	US electronic healthcare data	115

diagnosis code within an inpatient visit. The third phenotype definitions were based on a narrower set of diagnosis codes derived from ICD-based code lists. The fourth nonhemorrhagic stroke phenotype used a broader set of diagnosis codes. Full details of the definitions are available in Supplementary section A.

Right censoring can occur as we are using retrospectively collected observational data to develop the models using a cohort design. This is when patients are not observed for the full 1-year follow-up. In this study, we included patients who were only partially observed for the 1-year follow-up based on a prior study [16]. It was shown that a small amount of class label noise is better than introducing bias by removing patients with uncertain class labels.

Predictors

Candidate predictors were constructed using one-hot encoding for any medical code, drug code, procedure code, measurement code, or observation code recorded in the database. In addition, age in 5-year buckets (e.g., 0–4, 5–9, 10–14) and gender were also included. In total, over 19,000 candidate predictors were used in this study, but this number varied by database. These predictors have been used in prior studies and result in better-performing models than using a small number of prespecified predictors when developing LASSO logistic regression models [17].

Model development

We developed models for each combination of database, population, and outcome definition per prediction task. For each model, we trained a LASSO logistic regression model [18] using threefold cross validation with the train data (75% of the data) to pick the optimal regularization hyper-parameter. Three-fold cross validation has been shown sufficient for big data [19]. Using the optimal hyper-parameter value, the hyper-parameter that maximized the likelihood, the final model was fit using all the train data. The test data (remaining 25% of the data) was used to internally validate the model.

Validation

Validation of each model was performed by applying the model to all test datasets for the same prediction task

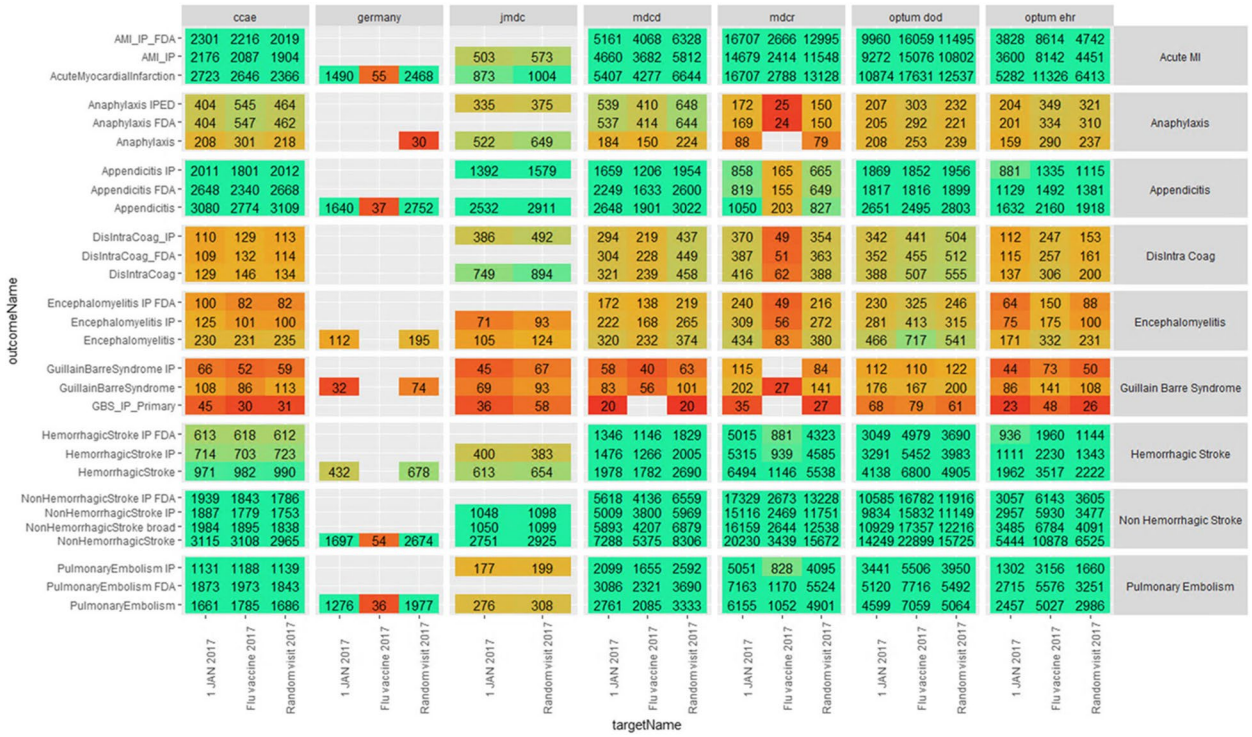

Fig. 1 Heatmap showing the number of outcomes in each dataset used to develop the models. Each cell corresponds to a dataset created with a database, population, and outcome definition combination. The cell values correspond to the number of patients in the dataset with the outcome during the time at risk. Approximately half of the datasets had > = 1000 patients with the outcome. IP, inpatient visit; FDA, Food and Drug Administration; IPED, inpatient or emergency department visit

(permuting databases, populations, and/or outcome definitions). This corresponded to each model being validated > 50 times (7 databases × 3 populations × 3 outcome definitions).

Performance metrics

Performance was evaluated using the area under the receiver operating curve (AUROC) [20], area under the precision recall curve (AUPRC), scaled Brier score, and calibration-in-the-large ratio (mean predicted risk divided by mean observed risk). The AUROC is a measure of discrimination that corresponds to the probability that a randomly selected patient who had the outcome in the year after index will be assigned a higher risk by the model than a randomly select patient who did not have the outcome. An AUROC of 0.5 corresponds to randomly assigning risk, and an AUROC of 1 corresponds to a model that can perfectly discriminate between those who will experience the outcome vs those who will not. AUPRC is another measure of discrimination but is preferred over the AUROC when the outcome is rare. The scaled Brier score is a measure of prediction accuracy. It corresponds to 1 minus the Brier score (mean squared error) divided by the Brier score for a model that predicts the mean observed risk. The calibration-in-the-large ratio provides insight into the model calibration on average.

Stability performance

To assess performance stability (level 1), we applied all models developed for the task of interest (e.g., nonhemorrhagic stroke) to each test dataset for this task (all combinations of database, population, and outcome definition). We then compare the performance of the model developed using the test dataset (internal performance) with the "stability" performances of all the other models developed with different datasets for the task. We then plot the internal performance against the stability performances to visualize how stable the performances are across choice of database, population, and outcome definition.

To assess individual risk stability (level 4) we calculated the mean absolute predictor error (MAPE). The per patient MAPE is calculated as the mean absolute difference between the original model (model developed on the same dataset that the patient is from) predicted risk for the patient and the sensitivity models' (models developed on data where the database, population, and/or outcome definition was permuted) predicted risks for the patient. We then calculate the average MAPE across all patients in the test set. We plot the original model prediction against other predictions obtained from the sensitivity models (prediction instability plot) and the original model prediction against the MAPE (MAPE instability plot).

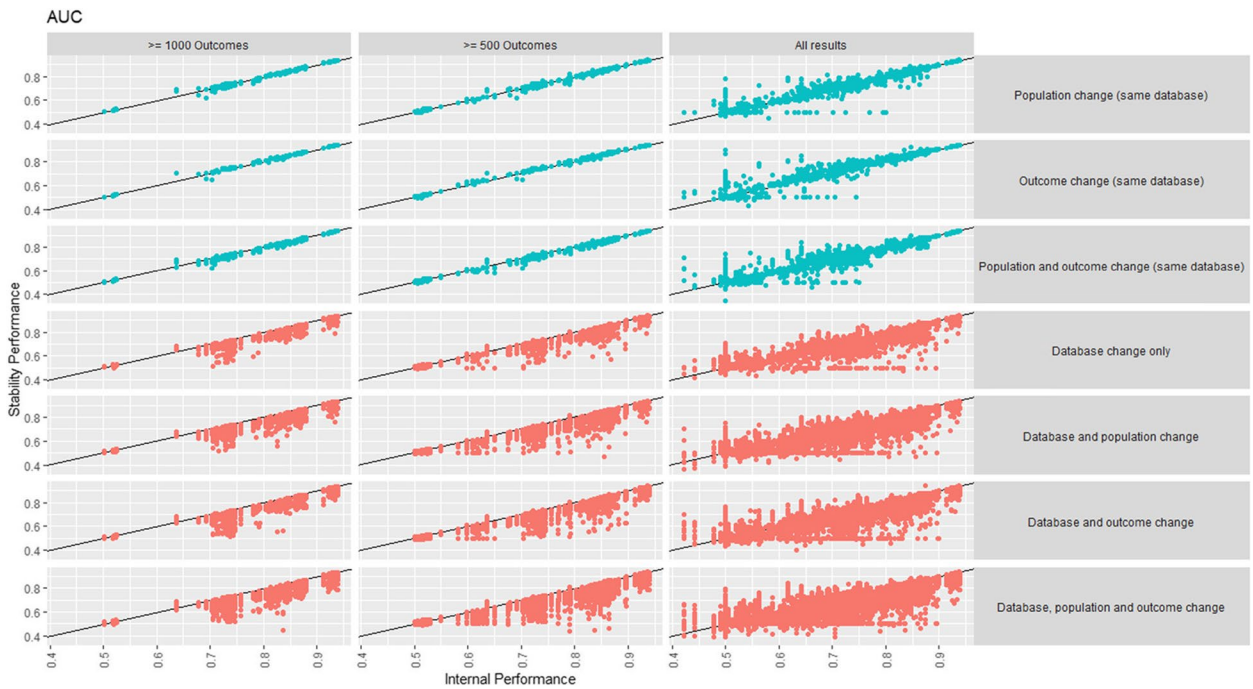

Fig. 2 Internal AUROC vs stability AUROC when the models are developed and validated in the same database (top three rows) or different database (bottom four rows). The columns "> = 1000 Outcomes," "> = 500 Outcomes," and "All results" represent only including models and validations where the dataset contained > = 1000 patients with the outcome, only including models and validations where the dataset contained > = 500 patients with the outcome, and all models and validations, respectively. Each row corresponds to a different validation data permutation compared to the model development data (e.g., the population, outcome definition, and/or the database is changed)

The code to perform the analyses is available at https://github.com/ohdsi-studies/Covid19VaccinePrediction and readily implementable on any data mapped to the OMOP common data model.

Results and discussion

The number of outcomes within the population varied by database and outcome definition (see Fig. 1). Most datasets contain approximately 2 million patients, see Supplementary B: Fig. 1. Guillain–Barre syndrome and encephalomyelitis were the rarest outcomes investigated, with many databases containing less than 100 patients with the outcome. The cardiovascular outcomes were generally the most common with most databases and outcome definitions containing more than 1000 patients with the outcome, which is the number that the LASSO logistic regression performance often starts to stabilize [21]. Four-hundred seventy-five models were developed, and internal discriminative performance across these models varied (see Supplement C). Comparing train AUROC and test AUROC indicates most models were not overfit (see Supplement C). A total of 237 (49.9%) models were developed using data with >1000 patients with the outcome. We focus on results corresponding to the 237 models (and validations in the 237 datasets with > = 1000 patient with the outcome) as these models were less likely to be overfit and the validation performance point estimates will be more stable.

Figures 2, 3 and 4 show the AUROC stability, calibration stability, and Brier score stability, respectively, when the population, outcome definition, or both are changed between model development and validation. The bottom four rows correspond to when the model development and validation had a different database (colored red), and the top three rows correspond to when the model development and validation were on the same database (colored blue). AUPRC showed a similar trend to AUROC (see Supplement D).

Figures 5 and 6 illustrate the individual prediction risk stability. Figure 5 shows that overall changing the population, outcome definition, and/or database can result in very unstable individual prediction risk estimates as there was large variance in the predicted risks per patient.

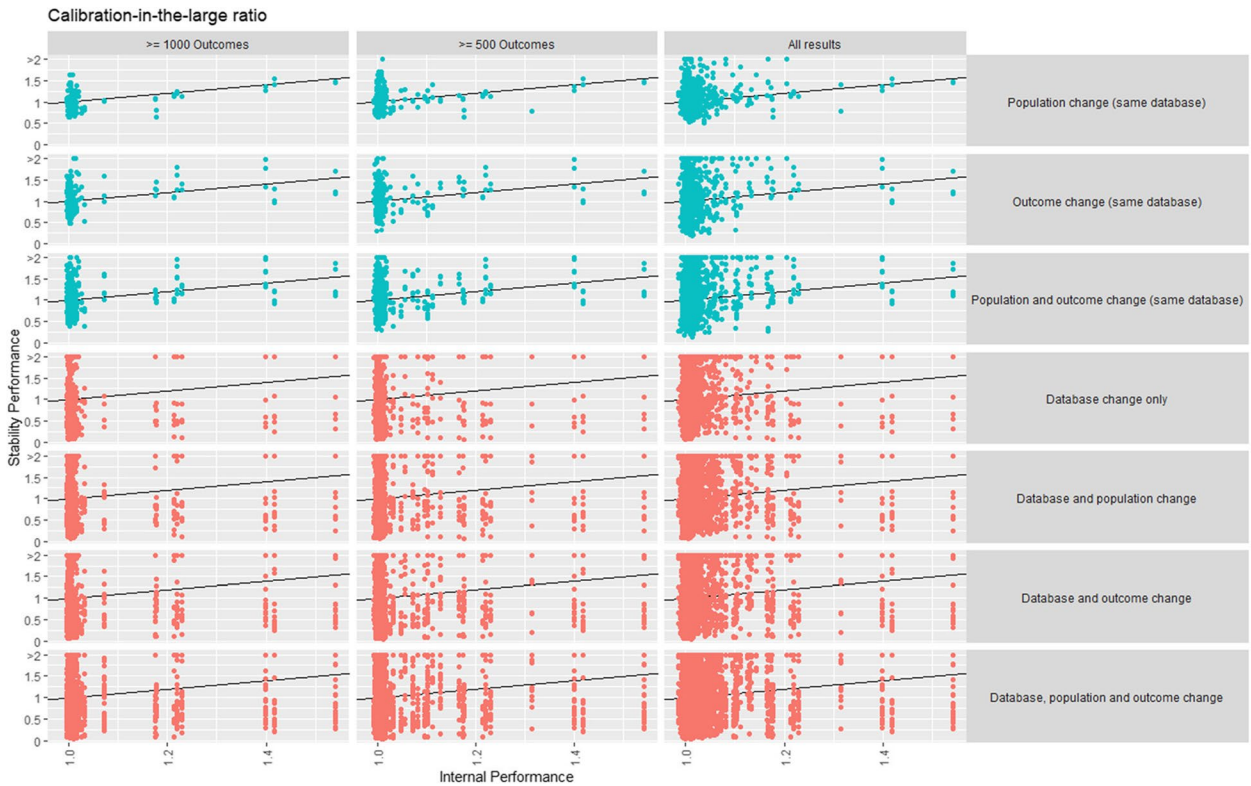

Fig. 3 Internal calibration-in-the-large ratio (mean predicted risk divided by mean observed risk) vs stability calibration-in-the-large ratio when the models are developed and validated in the same database (top three rows) or different database (bottom four rows). The columns "> = 1000 Outcomes," "> = 500 Outcomes," and "All results" represent only including models and validations where the dataset contained > = 1000 patients with the outcome, only including models and validations where the dataset contained > = 500 patients with the outcome, and all models and validations, respectively. Each row corresponds to a different validation data permutation compared to the model development data (e.g., the population, outcome definition, and/or the database is changed)

Figure 6 shows the MAPE is generally higher when the development and validation databases differed, but there was still instability in individual risk estimates when only the population or outcome definition differed.

Changing population

Figure 2 shows that the AUROC was generally stable when only the population was changed as the row corresponding to "Population change (same database)" shows the dots are on or near the x = y line when the database was consistent. However, changing the population and database can result in unstable AUROC (the dots on the row "Database and population change" were sometimes far from the x = y line). The AUPRC had a similar trend to the AUROC (see Supplement D). Interestingly, the scaled Brier score, the calibration-in-the-large ratio, and the individual predicted risks were unstable when only the population changed and even more unstable when the population and database changed.

The results suggest that discrimination, which often focuses on how well the model ranks patients based on risk, is only moderately impacted when there is only a change in population between model development and validation. However, the actual predicted risk value is unstable. It may be possible to improve predicted risk value stability by recalibrating the model in the dataset the model is transported into.

Changing outcome definition

Figure 2 shows that the AUROC was generally stable when only the outcome definition was changed as the dots fall on or near the x = y line for the row "Outcome change (same database)." However, changing the outcome definition and database can result in unstable AUROC (as the dots were sometimes far from the x = y line for the row "Database and outcome change"). The AUPRC had a similar trend to the AUROC (see Supplement D). Interestingly, the scaled Brier score, the calibration-in-the-large ratio, and the individual predicted risks were unstable when only the outcome changed and even more unstable when the outcome and database changed. Changing the outcome definition had a similar impact to changing the population.

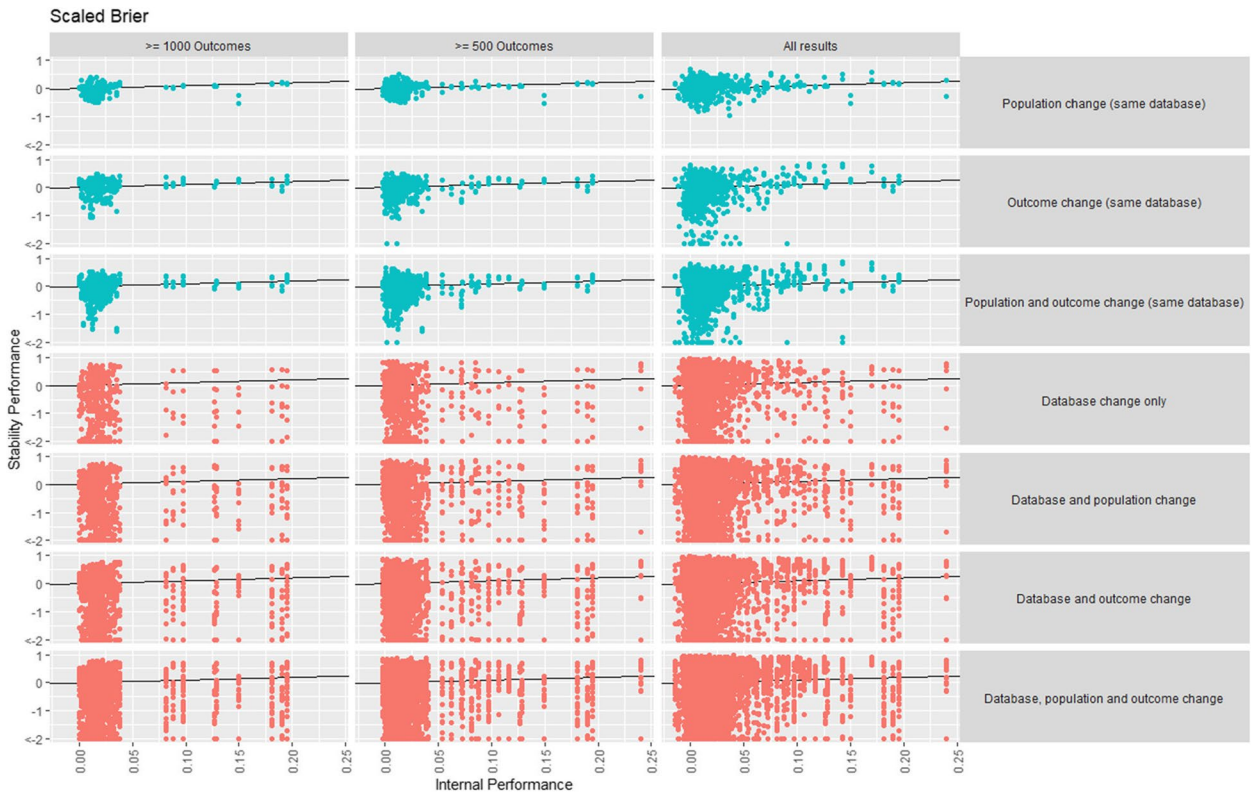

Fig. 4 Internal scaled Brier score vs stability scaled Brier score when the models are developed and validated in the same database (top three rows) or different database (bottom four rows). The columns "> = 1000 Outcomes," "> = 500 Outcomes," and "All results" represent only including models and validations where the dataset contained > = 1000 patients with the outcome, only including models and validations where the dataset contained > = 500 patients with the outcome, and all models and validations, respectively. Each row corresponds to a different validation data permutation compared to the model development data (e.g., the population, outcome definition, and/or the database is changed)

Changing database

Figures 3, 4, and 5 show that when the database was changed, all performance metrics were unstable as the red dots did not fall on or near the x = y line for the rows indicating the database was changed. Changing the database appears to have the greatest impact on model performance stability across all metrics and individual risk predictions.

Changing population, outcome definition, and/or database

AUROC and AUPRC were stable across changing the population and outcome definition. The accuracy metrics such as scaled Brier score and calibration-in-the-large were unstable. Individual risks were also unstable. Any change that included changing the database led to highly unstable performance metric results and individual predicted risk results.

In summary, we see that models are unstable, both in terms of population-level performance and individual-level predicted risk, when implemented in a new database. If only the population and outcome definitions are changed, the discriminative performance is stable, but other metrics and individual risks are unstable. Given this information, we highlight the importance of targeted validation, where a model developed in observational data is tested in the clinical setting it will be implemented in prior to use, as the model may perform much worse, which potentially could cause harm. In addition, recalibration in the target setting may be required.

Limitations

For the population sensitivity, the index date differed (Jan 1 st, random visit, influenza date), but some patients may overlap between the populations when the model was developed and validated in the same database. Therefore, there may be correlations between the patients used to develop the model and the patients used to validate the model when assessing stability. This may make the models appear more stable.

In this study we only investigated three population cohorts. The results may not hold for different populations. In addition, we only considered 3–4 outcome definitions per prediction task, and the definitions used were developed carefully. The result that the discrimination performances are often stable across slight changes to the outcome definition may not hold if a highly inaccurate outcome definition is used to develop a model.

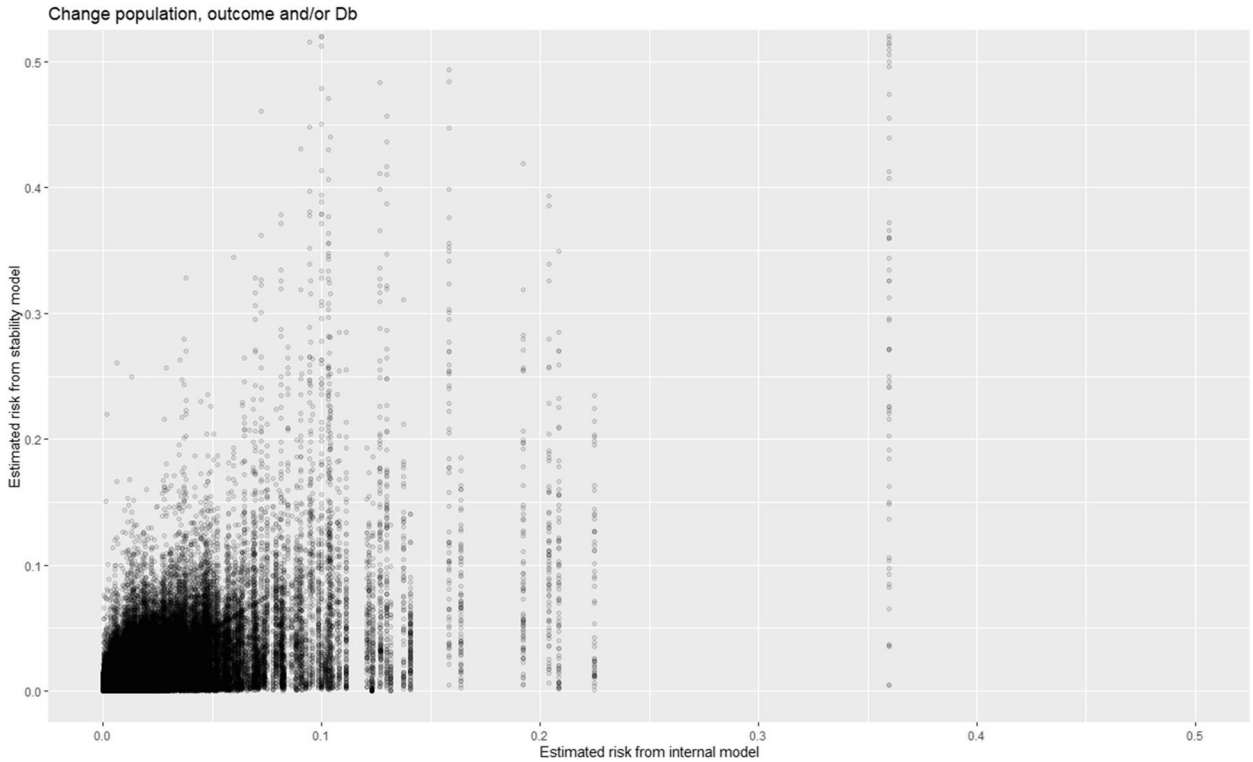

Fig. 5 Prediction instability plot showing the stability predictions (of models developed using a different population, outcome definition, and/or database) against the internal prediction for a random selection of 100 patients per test set

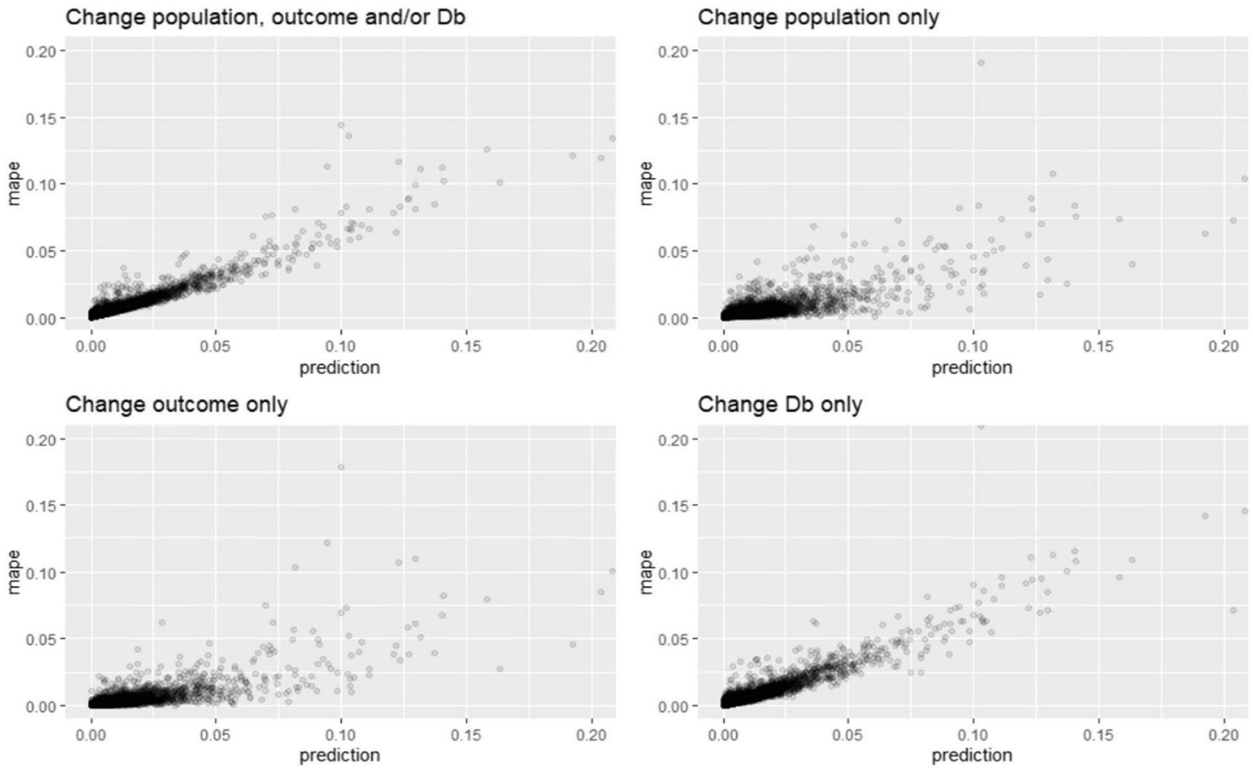

Fig. 6 MAPE instability plot when applying models that (1) have a change in population, outcome, and/or database, (ii) models that only have a change in population between development and validation, (iii) models that only have a change in outcome definition between development and validation, and (iv) models that only have a change in database between development and validation. These plots used a randomly selected 100 patients from each test set

The results show that the performances were more unstable when including models developed using data with a low number of patients with the outcome and including validations in data where there was low number of patients with the outcome. It is unclear whether this is due to unstable point estimates or unstable models. However, we presented models and validations that used data with >1000 patients with the outcome separately to try and minimize the impact of model overfitting and point estimate instability on stability insights.

In this study, we developed generalized linear models that did not include interaction terms. It is possible to develop models that work well across case mixes by adding suitable interaction terms. Future work could replicate this study but include age/sex interaction terms to see whether the performance and individual risks are more stable across databases when interactions are used.

Finally, we did not perform any form of recalibration using the validation data. Miscalibration could be fixed by recalibrating using some of the validation data. This may lead to more stable Brier score, calibration, and individual risk estimates. Future work could investigate recalibrating the models to see whether that improves stability.

Conclusion

This study investigated the impact that changing the database or population or outcome definition has on prediction performance and individual risk estimates. This is important if researchers are aiming to develop models using a single database that would be applied outside the database setting (e.g., in a real-world clinical population). The results show that small changes in the outcome definition are unlikely to have a large impact on the discriminative performance. This is important as outcome phenotype definitions are likely to have measurement error in observational data. Surprisingly, the discrimination performances were generally robust across different populations and index dates in this study (first visit, influenza vaccine visit, or random visit) but were unstable when the database changed. Calibration and Brier score appear to be unstable across changes to the outcome definition, population, and database. This highlights the need to recalibrate models when they are transported into new patient populations.

Model performance was also more unstable when the outcome is rare (< 1000 patients with the outcomes in the data). We therefore recommend, in agreement with other publications [8], that researchers validate any model in the clinical setting it will be applied to prior to using it for decision-making. In addition, we suggest a performance sensitivity analysis is implemented investigating different population definitions and outcome definitions if a model is developed in a database where the outcome occurs in less than 1000 patients (i.e., the outcome is rare).

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s41512-025-00191-x.

Supplementary Material 1: Supplementary A: Target 1: 1 st Jan 2017. Target 2: Flu vaccine 2017. Target 3: Random visit 2017. AMI Outcome 1: Acute Myocardia Infarcion. AMI Outcome 2: AMI IP. AMI Outcome 3: AMI IP_FDA. Guillain Barre Syndrome Outcome 1: Guillain Barre Syndrome. Guillain Barre Syndrome Outcome 2: Guillain Barre Syndrome IP. Guillain Barre Syndrome Outcome 3: Guillain Barre Syndrome IP primary events. Encephalomyelitis Outcome 1: Guillain Barre Syndrome. Encephalomyelitis Outcome 2: Guillain Barre Syndrome IP. Encephalomyelitis Outcome 3: Guillain Barre Syndrome IP FDA. isseminated intravascular coagulation Outcome 1: DisIntraCoag. Disseminated intravascular coagulation Outcome 1: DisIntraCoag. Disseminated intravascular coagulation Outcome 2: DisIntraCoag IP. Disseminated intravascular coagulation Outcome 2: DisIntraCoag IP. Anaphylaxis Outcome 1: Anaphylaxis. Anaphylaxis Outcome 2: Anaphylaxis IPED. Anaphylaxis Outcome 3: Anaphylaxis IPED FDA. Appendicitis Outcome 1: appendicitis. Appendicitis Outcome 2: appendicitis IP. Appendicitis Outcome 3: appendicitis IPED. Appendicitis Outcome 3: appendicitis IPED FDA. Hemorrhagic Stroke Outcome 1: Hemorrhagic Stroke. Hemorrhagic Stroke Outcome 2: Hemorrhagic Stroke IP. Hemorrhagic Stroke Outcome 3: Hemorrhagic Stroke IP FDA. Non-hemorrhagic Stroke Outcome 1: non-hemorrhagic Stroke broad. Non-hemorrhagic Stroke Outcome 2: Non-hemorrhagic Stroke broad IP. Non-hemorrhagic Stroke Outcome 3: Non-hemorrhagic Stroke broad IP FDA. Pulmonary Embolism Outcome 1: Pulmonary Embolism. Pulmonary Embolism Outcome 2: Pulmonary Embolism IP. Pulmonary Embolism Outcome 3: Pulmonary Embolism FDA. Supplementary B: Dataset sizes. Figure 1- Dataset size in thousands. Supplementary C: Internal Validation. Figure 2—AUROC on the test set across models. Figure 3—AUROC on the train set across models. Supplement D: Results for all models. Figure 4 internal AUPRC vs stability AUPRC when the models are developed and validated in the same database (top three rows) or different database (bottom four rows). The columns '> = 1000 Outcomes', '> = 500 Outcomes' and 'All results' represent only including models and validations where the dataset contained > = 1000 patients with the outcome, only including models and validations where the dataset contained > = 500 patients with the outcome and all models and validations, respectively. Each row corresponds to a different validation data permutation compared to the model development data (e.g., the population, outcome definition and/or the database is changed).

Authors' contributions

JMR, PBR and PRR designed the analyses. JMR performed the analysis and was a major contributor in writing the manuscript. All authors reviewed and approved the final manuscript.

Funding

This study received no funding.

Data availability

Data may be obtained from a third party and are not publicly available. The MarketScan Commercial Claims, MDCD and MDCR data that support the findings of this study are available from Merative (contact at: https://www.merative.com/documents/brief/marketscan-explainer-general) and the Optum EHR and CDM datasets are available from Optum (contact at https://www.optum.com/en/business/life-sciences/real-world-data.html), but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Declarations

Ethics approval and consent to participate

The use of the anonymized data in this study was reviewed by the New England Institutional Review Board and was determined to be exempt from broad Institutional Review Board approval.

Consent for publication

Not applicable.

Competing interests

JMR and PBR are employees and shareholders of Johnson & Johnson. PRR works for a research group that received/receives unconditional research grants from Yamanouchi, Pfizer-Boehringer Ingelheim, Novartis, GSK, Chiesi, Astra-Zeneca, Amgen, Janssen Research & Development.

Received: 24 July 2024 Accepted: 9 April 2025
Published online: 17 April 2025

References

1. Clift AK, Coupland CA, Keogh RH, Diaz-Ordaz K, Williamson E, Harrison EM, Hayward A, Hemingway H, Horby P, Mehta N, Benger J. Living risk prediction algorithm (QCOVID) for risk of hospital admission and mortality from coronavirus 19 in adults: national derivation and validation cohort study. Bmj. 2020;371:m3731.
2. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335(7611):136.
3. Lee TH, Marcantonio ER, Mangione CM, Thomas EJ, Polanczyk CA, Cook EF, Sugarbaker DJ, Donaldson MC, Poss R, Ho KK, Ludwig LE. Derivation and prospective validation of a simple index for prediction of cardiac risk of major noncardiac surgery. Circulation. 1999;100(10):1043–9.
4. Wynants L, Van Calster B, Collins GS, Riley RD, Heinze G, Schuit E, Albu E, Arshi B, Bellou V, Bonten MM, Dahly DL. Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal. bmj. 2020;7:369.
5. Williams RD, Markus AF, Yang C, Duarte-Salles T, DuVall SL, Falconer T, Jonnagaddala J, Kim C, Rho Y, Williams AE, Machado AA. Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network. BMC Med Res Methodol. 2022;22(1):1–13.
6. Reps JM, Kim C, Williams RD, Markus AF, Yang C, Duarte-Salles T, Falconer T, Jonnagaddala J, Williams A, Fernández-Bertolín S, DuVall SL. Implementation of the COVID-19 vulnerability index across an international network of health care data sets: collaborative external validation study. JMIR Med Inform. 2021;9(4):e21547.
7. Patel J, Scott F, Mohan R. It's a risky business: use of the QCovid risk calculator in a psychiatric rehabilitation population to enhance prevention. BJPsych open. 2021;7(S1):S46–S46.
8. Sperrin M, Riley RD, Collins GS, Martin GP. Targeted validation: validating clinical prediction models in their intended population and setting. Diagnostic and prognostic research. 2022;6(1):24.
9. Wessler BS, Nelson J, Park JG, McGinnes H, Gulati G, Brazil R, Van Calster B, van Klaveren D, Venema E, Steyerberg E, Paulus JK. External validations of cardiovascular clinical prediction models: a large-scale review of the literature. Circulation: Cardiovascular Quality and Outcomes. 2021;14(8): e007858.
10. Shivade C, et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. Journal of the American Medical Informatics Association. 2014;21(2):221–30.

11. Cai X, Li Y. Are AMI patients with comorbid mental illness more likely to be admitted to hospitals with lower quality of AMI care? PLoS ONE. 2013;8(4): e60258.
12. Mentz RJ, et al. Assessment of administrative data to identify acute myocardial infarction in electronic health records. Journal of the American College of Cardiology. 2016;67(20):2441–2.
13. Riley RD, Collins GS. Stability of clinical prediction models developed using statistical or machine learning methods. Biom J. 2023;65(8): 2200302.
14. Voss EA, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. Journal of the American Medical Informatics Association. 2015;22(3):553–64.
15. Reps JM, et al. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. Journal of the American Medical Informatics Association. 2018;25(8):969–75.
16. 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 Med Inform Decis Mak. 2021;21:1–24.
17. Reps JM, Wong J, Fridgeirsson EA, Kim C, John LH, Williams RD, Fisher RR, Ryan PB. Finding a constrained number of predictor phenotypes for multiple outcome prediction. BMJ Health & Care Informatics. 2025;32(1):e101227.
18. Suchard MA, Simpson SE, Zorych I, et al. Massive parallelization of serial inference algorithms for complex generalized linear models. ACM Transact Model Comput Simulation. 2013;231:10–32.
19. 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.
20. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982;143:29–36.
21. 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 Informatics. 2022;163: 104762.

Publisher's Note