BMJ Open

Does prognostic model development and internal validation design matter with big data?

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-050146
Article Type:	Original research
Date Submitted by the Author:	11-Feb-2021
Complete List of Authors:	Reps, Jenna; Janssen Research and Development LLC; Observational Health Data Sciences and Informatics Community Ryan, Patrick; Janssen Research and Development LLC; Observational Health Data Sciences and Informatics Community Rijnbeek, P; Erasmus University Rotterdam; Observational Health Data Sciences and Informatics Community
Keywords:	PREVENTIVE MEDICINE, STATISTICS & RESEARCH METHODS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Does prognostic model development and internal validation design matter with big data?

Jenna M. Reps, PhD[1,2*], Patrick B. Ryan[1,2], Peter R. Rijnbeek[1,3]

[1]Observational Health Data Sciences and Informatics Community, New York, NY, USA; [2]Janssen Research and Development, Raritan, NJ, USA; [3]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

***Corresponding author - Email: jreps@its.jnj.com**

Abstract

Objective: Internal validation of prediction models aims to quantify the generalizability of a model. We aim to determine the impact, if any, that the choice of internal validation design has on the internal discrimination estimate bias and model transportability in big data (n ~ 500,000).

Methods: Model discrimination were assessed using the area under the receiver operating curve (AUROC). We assess the impact of the development design across 21 real-world prediction questions. We trained LASSO logistic regression models using US claims data and internally validated the models using eight different designs: use test set (yes/no) and use validation set (no/cross validation with 3-,5- or 10-folds). We then externally validated each model in a new US claims database. We estimate the internal validation bias per design by empirically comparing the estimated internal discriminative performance and external performance.

Results: The differences between internal estimated AUROC and external AUROC was largest for the design that did not use a test set or cross validation (CV). Internal estimates of AUROC were > 0.1 more than the external AUROC across designs when the outcome event count was < 1000 (indicating biased estimates). Internal AUROC estimates were closer to external AUROCs when the outcome event count was >= 5000 for designs that used CV (indicating unbiased estimates). Across multiple prediction questions, CV only designs lead to more transportable models. However, the transportability of the models developed using a CV only design and CV with test set design were comparable when the outcome event count was >= 5000.

Conclusions: The choice of internal validation design impacts the amount of bias in the internal validation performance estimate even in big data unless the outcome event count is high (>5000). When the outcome count is < 1000 all designs can lead to optimistic internal discriminative estimates.

Strengths and limitations of this study:

- We developed and externally validated 840 prediction models using 8 different development/internal validation designs across 21 prediction problems
- We focused on a target population of approximately 500,000 patients and predicted 21 different outcomes of various rareness
- We empirically investigated the impact of development/internal validation design on internal discrimination estimate bias and model transportability in big data

Word count: 4138

Keywords:

patient-level prediction; internal validation; *sample size; best practices*

Background

Sub-optimal design choices for the model development and validation (e.g., how hyper-parameters are selected, how a classifier with specified hyper-parameters is trained and how a trained model's internal validation is estimated) could introduce model bias (i.e., an overfitted model or optimistic performance estimates) [1]. The idea of binary classification is to learn how to discriminate between two classes (e.g., healthy vs unhealthy or will develop cancer vs will be cancer free) using attributes/features/covariates e.g., (age, body mass index). In general, we can only get labelled data for a sample of the whole population, but we aim to develop models using the sample that can generalize to the whole population of interest. For example, when developing a prognostic model, we generally have n patients sampled from a bigger population, with p predictors (aka features/covariates) and outcome labels indicating whether the n patients had some outcome recorded during some time period of interest. When we develop prognostic models, the aim is to learn associations between the p predictors and the outcome label based on what we see from the n patients. However, we want the model to generalize such that the associations hold true for the population that the n patients were sampled from (e.g., a new sample of patients). When this is not the case, it is often referred to as model overfitting. When a model overfits we see excellent discrimination between the two classes in our population sample, but when the developed model is applied to the population that were not sampled it is not able to discriminate as well. When a prognostic model is developed, the same data are often used to estimate the 'internal validation' performance of the model. However, there are various methods that can be used to account for overfitting and calculate a less biased discriminative performance estimate, such as bootstrapping, cross-validation or using a held-out sample of data from the n patients (test set). 'External validation' is when a model is applied to data corresponding to new patients and the performance is estimated.

Studies on small data (n <10,000), with a few candidate predictors (p < 100) and using a simple logistic regression have shown that a bootstrapping design leads to developed models with the most accurate internal discrimination performance estimates [2-3]. Bootstrapping requires fitting models a large number of times, often hundreds of times, and this makes it unsuitable for big data. There is currently no empirical study comparing the impact of different model development designs when data are large (big n and big p) on the accuracy of the internal validation and the external validation.

When developing prognostic models that use binary classifiers the classifier has parameters that need to be determined based on the data and may also have hyper-parameters that need to be selected. For example, considering a regularized logistic regression, the parameters are the coefficients/intercept and the hyper-parameter is the amount of regularization. For a decision tree, the parameters are the rules that do the splitting, and the hyper-parameters are things such as the maximum depth and the minimum number of data points at each leaf. The hyper-parameters effectively determine how complex a decision boundary can be. Some binary classifiers such as Naïve Bayes and logistic regression (with no regularization) have no hyper-parameters and therefore only require learning the optimal parameter values. In this manuscript we are going to focus on one classifier, the LASSO logistic regression, that has a single hyper-parameter that controls the penalty assigned to model complexity.

There are two steps in training a binary classifier: i) learning the optimal hyper-parameters and ii) learning the parameters. If we split the data into just training data (data used to develop the model) and test data (data held-out for internal evaluation), then we would use all the training data to learn the

parameters for a set of candidate hyper-parameter settings and then we would apply all the models (for each hyper-parameter setting) to the test set and pick the model that did best. However, then we have used the test set to pick the model rather than as an independent data set that is just used to evaluate the model and we cannot estimate the performance fairly. As a consequence, the standard process for model development is to split the data into training/validation/test data. The train data are used to learn the model parameters, the validation data are used to assess the model hyper-parameters and the test set is used to fairly evaluate the final model (with the optimal hyper-parameter learned on the validation set).

Unless the data are very large, is it undesirable to split the data three times, as that introduces uncertainty around the performance estimates. To solve this, a common technique is to combine the training/validation data and use a process known as cross-validation (CV) to both learn the parameters and hyper-parameters on the same data but using a process that enables a fair evaluation of the hyper-parameters. N-fold CV works by partitioning the training/validation data into N disjoint sets, termed folds. Then, for each fold, the fold data is held out while the other folds are used to train the model. The trained model is then applied to the left-out fold to calculate the measure of performance. This is repeated for each fold, to give N performance estimates. The mean of the N estimates is then calculated and used as the CV estimated performance. It is possible to apply CV on all the data to estimate the internal validation as an alternative to using a test dataset. However, using CV on all the data only estimates the final model's performance based on the selected hyper-parameters as the actual final model is never truly evaluated on new data (i.e., the models' parameters for each fold may differ from the final model parameters).

Another technique for estimating the internal validation fairly is known as bootstrapping. When applying bootstrapping the model is trained using all the data and then the potential overfitting is quantified by repeating the training process multiple times (100s) using random samples drawn from the data with replacement and calculating the average different between the sample models' AUCs on the sample data and complete data. Bootstrapping would still require a process such as CV to determine the hyper-parameters. When the data are large, this method is not feasible as if model fitting took days, then repeating the process 100 times would take too long. When the data are small, this method has been shown to provide the fairest estimates of internal validation.

In this paper we compare the impact of model development design on the interval validation and external validation in big data. We focus on data with approximately 500,000 patients and investigate performance estimates across 21 prediction problems with varying outcome event count rareness. We implemented eight different development designs per prediction problem: combinations of whether to use a test set (yes/no) and whether to use cross validation (no, 3-fold, 5-fold or 10-fold). We repeated each design multiple times with different splits (folds or test sets) to estimate how stable the internal performance estimates are. We then investigate whether the designs lead to any differences in model performance when externally validating the models in a new database.

Methods

We use the OHDSI PatientLevelPrediction framework [4] and R package to develop and evaluate the prediction models in this study.

Data

We developed models using a US claims database IBM CCAE that contains insurance claims data for commercially employed individuals and their dependents. The patients in this database are aged under 65. The database contains records for approximately 153 million patients between Jan 2000 to Dec 2019.

Models were externally validated using IBM MDCR, a US claims database for Medicare patients with supplemental insurance. MDCR contains insurance claims records for patients mostly aged 65 or older. The database contains approximately 10 million patients and records from Jan 2000 to Jan 2020.

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

Patient and Public Involvement

No patient involved

Study Population

We extracted data for patients who are pharmaceutically treated for their first occurrence of depression to predict 21 outcomes occurring for the first time from 1 day after their depression diagnosis until 365 days after. In the development data we randomly sampled 500,000 patients from 1,964,494 treated for depression and this resulted in a range of outcome event count sizes during the 1-year follow-up. In the external validation data, we used all the data available, this corresponded to 160,956 patients.

Outcomes

We used the same 21 outcomes used by the PatientLevelPredicion framework study [4]. Table 1 lists the 21 outcomes we predicted occurring 1 day after index until 365 days after index. The number of outcome events in the development data and validation data are also reported. As we are predicting first occurrence of each outcome, we excluded patients with the outcome prior to their depression, so the study populations slightly differed per outcome (e.g., when predicting acute liver injury we exclude patients with a history of acute liver injury but when predicting ischemic stroke we exclude patients with a history of ischemic stroke).

Table 1- Outcomes predicted in this study and the logic used to define the outcome in the data.

Outcome	Phenotype	Event count in development data (N ~ 500,000)	Event count in validation data (N ~ 160,956)
Open angle glaucoma	A first-time condition record of Open-angle glaucoma with at least 1 condition record of Open-angle glaucoma from a provider with Ophthalmology, Optometry or Optician speciality within 1 to 365 days.	174	510
Acute liver injury	A first-time condition record of Acute liver injury during an ER visit or inpatient visit. No	184	67

	Acute liver injury exclusions 1 year prior to 60 days after.		
Ventricular arrhythmia and sudden cardiac death	A first-time condition record of Ventricular arrhythmia and sudden cardiac death during an ER visit or inpatient visit being the primary cause of the visit.	297	642
Ischemic stroke	A first-time condition record of Ischemic stroke during an inpatient visit	380	1153
Acute myocardial infarction	A first-time condition record of Acute myocardial infarction during an ER visit or inpatient visit being the primary cause of the visit.	491	1080
Gastrointestinal hemhorrage	A first-time condition record of Gastrointestinal hemhorrage during an ER visit or inpatient visit being the primary cause of the visit.	509	963
Delirium	A first-time condition record of Delirium during an ER visit or inpatient visit	985	1298
Seizure	A first-time condition record of Seizure during an ER visit or inpatient visit	1494	935
Decreased libido	A first-time condition record of Decreased libido	1661	130
Alopecia	A first-time condition record of Alopecia	2577	748
Hyponatremia	A first-time condition record of Hyponatremia or a first-time measurement of serum sodium between 1 and 136 millimole per litre	2628	4276
Fracture	A first-time condition record of Fracture	2722	4071
Vertigo	A first-time condition record of Vertigo	3046	2086
Tinnitus	A first-time condition record of Tinnitus	3120	1824
Hypotension	A first-time condition record of Hypotension	4170	6399
Hypothyroidism	A condition record of Hypothyroidism with another condition record of Hypothyroidism within 90 days	6117	3853
Suicide and suicidal ideation	A first-time condition record of Suicide and suicidal ideation or a first-time observation of Suicide and suicidal ideation	10221	993

Constipation	A first-time condition record of Constipation	10672	7569
Diarrhea	A first-time condition record of Diarrhea	14875	7226
Nausea	A first-time condition record of Nausea	19754	7824
Insomnia	A first-time condition record of Insomnia	20806	6846

Candidate Predictors

We used one-hot encoding for any medical event, drug, procedure, observation or measurement recorded within 1 year prior to, or on, index (date of depression). This means we have a binary predictor per medical event/drug/procedure/observation/measurement recorded for any patient in our study population within 1-year prior to index. For example, if a patient had a record of 'type 2 diabetes' 80 days prior to index, the value for the predictor 'type 2 diabetes 1-year prior' would be 1. If a patient never had type 2 diabetes recorded, their value for the predictor 'type 2 diabetes 1-year prior' would be 0. We also created one-hot encoded variables for any medical event, drug, procedure, observation or measurement recorded within 30 days prior to, or on, index. In addition, we added one-hot encoded variables for age in 5-year groups (0-4,5-9,…, 95-99), index month (for seasonality), ethnicity, race and gender. Finally, the number of visits in the prior 30 days was also used as a candidate predictor. This resulted in approximately 86,000 candidate predictors. In this paper we focus on the impact of study design on internal validation estimation and therefore do not present the final developed models.

Model Development Designs

We investigate developing and internally validating LASSO logistic regression models [5] using the designs in Table 1. LASSO logistic regression in a generalized linear model that adds a penalty term to penalize the inclusion of predictors that are only weakly associated to the class label. This effectively performs feature selection during model training.

We compare the estimated internal validation when developing models using:

- No split: the hyper-parameters, parameters and performance are determined using all the data. This has a high risk of overestimating the performance and is included as a worst case scenario.
- Test/train/validation split: the hyper-parameters are selected using the validation data, the parameters are selected using the training data and the performance is estimated using the test set. This is the quickest design apart from the no split.
- N-fold CV: CV on all the data is used the select the hyper-parameters and estimated the performance. Parameters are selected using all the data.
- N-fold CV with a test set: CV on the training data is used the select the hyper-parameters and parameters are selected using all the training data. Performance is estimated using the test set.

The designs are summarized in Table 2. We investigate the impact of the number of folds (N is 3, 5 or 10) when performing CV. All designs that use CV to select the optimal hyper-parameters used the same hyper-parameter grid search. The test/train splits were done stratified by outcome, so the % of people in the test/train data with the outcome were the same.

Table 2- the different designs compared in this study.

Design	CV	Test set?	Hyper-parameter selection	Parameter training	Internal validation
No split (test_0_cv_0)	0	No	Using all data	Using all data	Using all data
Test/Train/Validation (test_1_cv_0)	0	Yes	Using 10% validation data	Using 80% training data	Using 10% test data
3-fold CV on all data (test_0_cv_3)	3	No	Using 3-fold CV on all data	Using all data	Using 3-fold CV on all data
Test/Train with 3-fold CV (test_1_cv_3)	3	Yes	Using 3-fold CV on 80% training data	Using 80% training data	Using 20% test data
5-fold CV on all data (test_0_cv_5)	5	No	Using 5-fold CV on all data	Using all data	Using 5-fold CV on all data
Test/Train with 5-fold CV (test_1_cv_5)	5	Yes	Using 5-fold CV on 80% training data	Using 80% training data	Using 20% test data
10-fold CV on all data (test_0_cv_10)	10	No	Using 10-fold CV on all data	Using all data	Using 10-fold CV on all data
Test/Train with 10-fold CV (test_1_cv_10)	10	Yes	Using 10-fold CV on 80% training data	Using 80% training data	Using 20% test data

Evaluation of Models

The area under the receiver operating curve (AUROC) was used to evaluate the discriminative performance (how well it ranks based on predicted risk). The AUROC is a measure that ranges between 0 and 1, with values less than 0.5 corresponding to discrimination worse than random guessing risk (e.g., patients who will experience the outcome are generally assigned a lower risk than patients who will not experience the outcome), a value of 0.5 corresponding to randomly guessing the risk and values great than 0.5 corresponding to better than random guessing. The closer the AUROC is to 1, the better the discrimination.

For the AUROC estimated using N-fold CV we have N estimates of the AUROC (per fold). We calculate the 95% confidence interval (CI) using the formula mean – 1.96*standard deviation of the N estimates. For the test set AUROC we calculated the 95% CI using the standard deviation based on the Mann-Whitney statistic.

Are some development designs and ways to estimate internal discrimination more likely to lead to optimistic AUROC estimates compared to other designs? To investigate how accurate (unbiased) the

internal discriminative performance estimate is for a given model, we will implement the model externally on new data and compare the internal AUROC performance estimate with the external AUROC performance. The difference, internal AUROC - external AUROC, gives an indication of whether a model has overfit to the development dataset, where a value close to zero suggests an unbiased internal discriminative estimate and a value much greater than zero suggests a biased internal discriminative estimate.

Are some development designs and ways to estimate internal discrimination more likely to result in the development of more transportable models compared to other designs (i.e., the model does better in new data)? The default PatientLevelPrediction design is “Test/Train with 3-fold CV” (test_1_cv_3). To compare the impact of design on transportability of a model we calculate the mean different between the AUROC values of the models developed using each other design and the external AUROC values of the models developed using the default “Test/Train with 3-fold CV” design. A negative AUROC means the “Test/Train with 3-fold CV” design resulted in a higher AUROC than the alternative design and positive value means the AUROC value was higher when the alternative design was used to develop the model. A value of approximately zero means the models developed by the different designs performed equally when externally validated.

Results

The characteristics of the development and validation study populations are displayed in Table 3. The development study populations consisted of younger patients compared to the external validation dataset. The two databases appear to have the same study population gender ratio and similar prior mean observation (number of days a patient has been active in the database prior to index).

Table 3 - the characteristics of the study populations

	Development Data (N ~ 500,000)	*Validation Data (N ~* 160,956)
Mean Age in years (sd)	*40 (15)*	*75 (7.8)*
Male Gender %	*31%*	*32%*
Mean days prior observation (sd)	*1474 (1205)*	*1585 (1192)*
Condition recorded in prior year (% of patients)		
Neoplastic Disease	*21.1%*	*45.7%*
Pain	*60.1%*	*74.4%*
Anxiety	*41.3%*	*28.6%*
Respiratory tract infection	*15.9%*	*12.0%*
Dementia	*0.0%*	*0.9%*
Obesity	*10.5%*	*10.6%*
Diabetes Mellitus	*8.9%*	*27.0%*
Hypertensive disorder	*24.7%*	*69.0%*
Heart disease	*9.2%*	*46.5%*
Hyperlipidaemia	*23.3%*	*56.3%*

Figure 1 displays the results of the AUROC values and 95% CI across designs for five reputations of using a test set internal validation design (in blue) and using a CV internal validation (in red). The rows correspond to the number of folds used by cross-validation and the columns correspond to the 21 different outcomes. The rarest outcomes are on the left and the most common are on the right. The performance when cross-validation was not used to select hyper-parameters is the top row (0 folds). Red represents the AUC performances for designs where CV or all the data are used to estimate the internal performance, blue represents the AUC performances of designs where a test set is used to estimate the

internal performance and black represents the external validation for each design. The top row (no CV) differs from the rows 2 to 4, where we see that picking a model hyper-parameter and parameters without CV or a test set lead to highly overfit models. The AUROC performance varied across the outcomes. In general, the external validation (black cross) was lower than all internal validation estimates, except for three outcomes (decreased libido, alopecia and Hypothyroidism). The internal validation estimates using a test set vs CV appear to be similar across outcomes and the external validation performances were generally equivalent across internal validation designs. The number of folds used in CV (3, 5 or 10) does not appear to impact the internal or external validation estimates, except for rare outcomes where the CIs are wider.

To investigate whether some designs are more likely to lead to optimistic internal discriminative estimates we calculated the difference between the internal validation performance and the external validation performance for each model. Figure 2 shows box plots for the difference between the internal AUROC and the external AUROC on the x-axis with the y-axis representing the design used to develop the model and estimate the internal AUROC for a range of outcomes. In addition to box plots for all the outcomes, the outcomes are partitioned into those with outcome counts < 1000 (rarer outcomes), outcomes with counts between 1000 and 5000 and outcomes with counts >= 5000 (common outcomes) to investigate the trends when the outcome count increases. The results show that i) all designs except the design without a test set or CV had a similar median and interquartile range for the differences between their internal and external estimated AUROCs and ii) the median difference between the internal and external AUROC estimates decreased as the outcome count increased. The median difference was greater than 0.1 for all designs when the outcome count was less than 1000 but the median differences was less than 0.05 for all designs (except no test and no CV) when the outcome count was greater than 5000.

Investigating whether certain designs can lead to more transportable models, we investigated the differences in external AUROC between the "Test/Train with 3-fold CV" (test_1_cv_3) design and each other designs for the different outcomes and repetitions. The results are displayed in Figure 3. The box plots show that the designs that included CV transported better than the designs that did not. Not performing CV resulted in a median decrease in external AUROC of 0.01 when a test set was used and a median decrease in AUROC of 0.07 when no test set was used compared to the design that used a test set and 3-fold CV. The number of folds used within CV by the designs that used a test set and CV had minimal impact on model transportability as the median AUROC differences between using 3-folds and 5- or 10-folds design was 0. The designs that used CV with no test set discriminated slightly better on average when externally validated, with a median AUROC increase of 0.002-0.003 across CV values (max increase of 0.024). The differences in external AUROC between the design with a test set plus CV and a design with CV but without a test set decreased as the outcome count increased with the median difference between these designs being approximately 0 when the outcome count was greater than 5000.

Discussion

We investigated the impact of LASSO logistic regression model development and internal validation design in big data (N ~ 500,000) for various outcomes. The results show that the design can impact whether an internal performance estimate is biased, even when developing models with big data. In general, the results show that using CV or a test set to estimate internal validation resulted in less biased internal AUROC estimates. However, evaluating a model internally using all the data without CV resulted in very biased AUROC estimates, even when the outcome count was greater than 20,000. Interestingly, the internal AUROC estimate became more similar to the external AUROC value for the designs using CV as the outcome count increased. Even with big target populations, if the outcome is rare (count less than 1000) then the internal AUROC estimate is likely to be optimistic. In terms of transportability, the models developed using CV without a test set were slightly better when externally validated (median improvement of 0.002-0.003 in AUROC) compared to the models developed using 3-fold CV and a test set. The gain in external validation AUROC by using a CV only design decreased as the outcome count increased. This makes sense, as the CV only design uses more data than the CV with a test set design, so when the outcome is rare, there is a gain in using more data (resulting in more outcomes to learn from) for model development. Therefore the choice of development and internal validation design used should focus on the number of outcomes in addition to the target population size.

This study has shown that developing LASSO logistic regression models using all available data and estimating internal validation via CV appears to result in slightly more transportable models when the outcome count is less than 5000. However, the improvement was generally moderate and using a test set has alternative advantages such as i) using less data to train will make training quicker and ii) a holdout set makes it possible to fairly evaluate variable importance. If the aim of the prediction is purely performance and the outcome count is less than 5000, then using a CV design on all the data may be preferable, however, if the interpretability of the model is important or the outcome count is greater than 5000, then using a CV design with a test set may still be preferable. Both CV using all data and CV with a test set had comparable internal validation accuracy. Interestingly, the number of folds used within CV made little impact on internal AUROC estimate accuracy nor transportability of the models. Therefore, it may be preferable to use 3-fold CV when data are large, as this will decrease the time it takes to develop the model.

Inspecting the performances of the designs per outcome, we see the external AUROC was lower for 17 out of the 21 outcomes. In about half of those cases, the external validation was off by more than 0.1. This is a large drop in performance, although not unexpected due to the external validation databases containing an older population. The results highlight the importance of performing external validation. If a poor internal validation design is used, but there is a large number of datasets used for external validation, then this is likely to give a more complete picture of the model's true ability across a range of patient populations compared to relying on just a single well-defined internal validation design with no external validation. For example, the design where we used all the data to pick the hyper-parameters, the parameters and estimate the interval performance was shown to consistently perform poorly when it was externally validated, even though the internal estimates indicated excellent discrimination.

The main strength of this study is that we were able to investigate the impact of development design across a large number of outcomes. In total we investigated 8 designs (with/without test set) and 3/5/10 CV or no CV, 21 outcomes and 5 repetitions, resulting in the development of 840 models (8x21x5). In addition, we externally validated each of these models. The percentage of the study population

experiencing each outcome ranged between ~0.04% to ~4%, enabling us to investigate the impact of development design when the outcome count was small and large.

We implemented a large-scale study to empirically estimate the impact that the model development design has on internal and external validation. However, there are some limitations of this study. Firstly, we only investigated developing models in one US claims data and in future work it would be useful to repeat this study using more datasets to see whether the results hold. Similarly, we only used one study population, patients initially treated for depression, and future work should investigate whether the results hold across different study populations and outcomes. Finally, we have only investigated the impact of the model development design when developing a LASSO logistic regression. Our results may not generalize to all binary classifiers.

Conclusion

Our study is the first to investigate the impact of model development design on the accuracy of the internal discrimination estimate and external validation performance when using big data (n=500,000). We compared using i) all the data to develop and validate a model, ii) a test set with no CV, iii) using CV with a test set and iv) using CV only to estimate the internal discriminative performance across 21 prediction problems. In addition, we investigated the impact of the number of CV folds. The results showed that not using CV or a test set in the design leads to overfitted models that have unrealistically high internal discrimination estimates. When CV was used to pick the hyper-parameters whether a test set was used had little impact on the internal discrimination estimates nor the external validation performance. The number of folds used during CV (3, 5 or 10 folds) appeared to have no impact on internal or external discrimination. These results show that even in big data CV should be used when developing LASSO logistic regression models and the choice of whether to use a test set depends on the purpose of the model and the outcome count.

Funding statement

This work was supported by the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Competing interests statement

Dr. Reps and Dr. Ryan report and are employees of Janssen Research and Development and are shareholders of Johnson & Johnson

Dr. Rijnbeek reports grants from Innovative Medicines Initiative, grants from Janssen Research and development, during the conduct of the study.

Author's contribution

JMR and PBR contributed to the conception and design of the work, JMR implemented the analysis. All authors contributed to the interpretation of data for the work and in drafting, revising and approving the final version.

Data sharing statement

The IBM CCAE and IBM MDCR data that support the findings of this study are available from IBM MarketScan Research Databases (contact at: http://www.ibm.com/us-en/marketplace/marketscan-researchdatabases) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

References

1. Wolff, R.F., Moons, K.G., Riley, R.D., Whiting, P.F., Westwood, M., Collins, G.S., Reitsma, J.B., Kleijnen, J. and Mallett, S., 2019. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Annals of internal medicine*, *170*(1), pp.51-58.
2. Steyerberg, E.W., Harrell Jr, F.E., Borsboom, G.J., Eijkemans, M.J.C., Vergouwe, Y. and Habbema, J.D.F., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, *54*(8), pp.774-781.
3. Steyerberg, E.W. and Harrell, F.E., 2016. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology*, *69*, pp.245-247.
4. Reps JM, 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):969-75.
5. Suchard MA, Simpson SE, Zorych I, et al. Massive parallelization of serial inference algorithms for complex generalized linear models. *ACM Transactions on Modeling and Computer Simulation*, 2013;23(1):10-32.

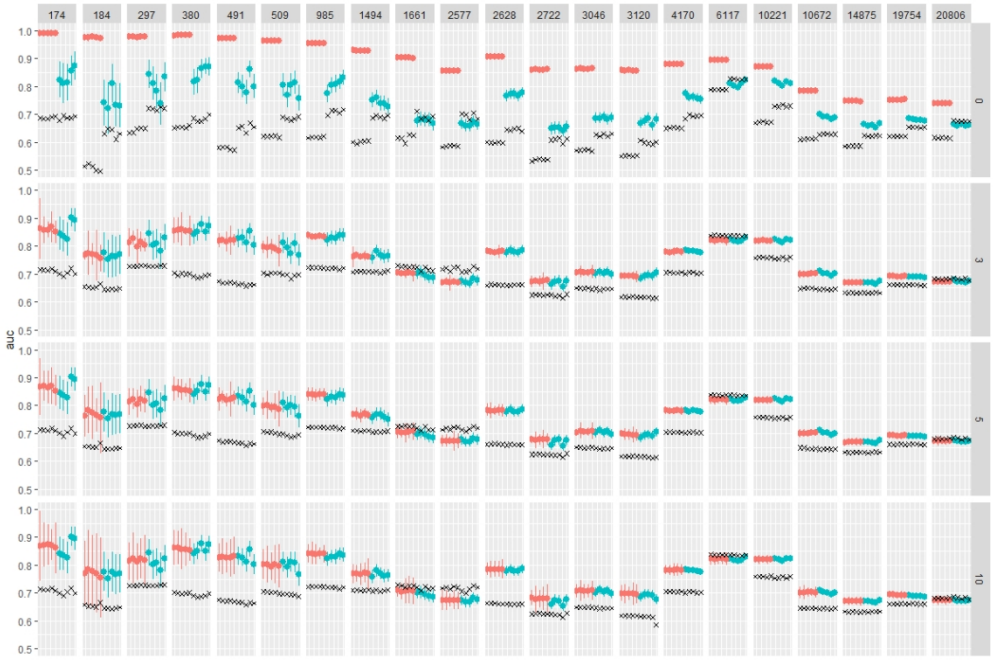


Figure 1: Comparison of the internal and external discrimination estimates across outcomes and designs. The x-axis corresponds to the models developed per design with and without a test set (5 repetitions using a test set and 5 repetitions not using a test set per outcome). Each dot is an estimated value for the model's AUROC discrimination, the line is the 95% CI and each black cross at the same x-value is the external validation AUROC discrimination for the model. The dot's color represents whether a test set was used (blue = test set, red = no test set). The columns are partitioned by outcome ordered by rareness (the number displayed is the number of outcome events) and the rows are the number of CV folds used by the design (0 - means no CV, or 3/5/10 folds).

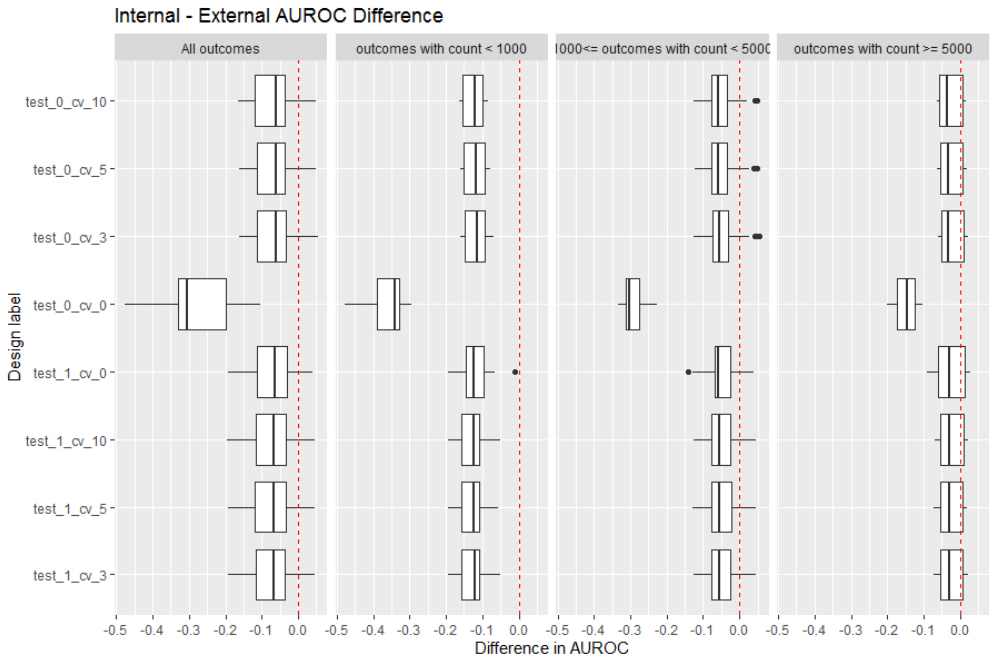


Figure 2: Box plots showing the Internal AUROC estimate minus the external AUROC estimate (x-axis) per design (y-axis). Values near 0 indicate that the internal validation AUROC estimates were accurate as the external validation AUROCs were similar. The first column contains box plots for all outcomes, and then remaining columns group outcomes into those with a count < 1000, those with a count between 1000 and 5000 and those with a count >= 5000. These enable the effect of outcome rareness to be inspected.

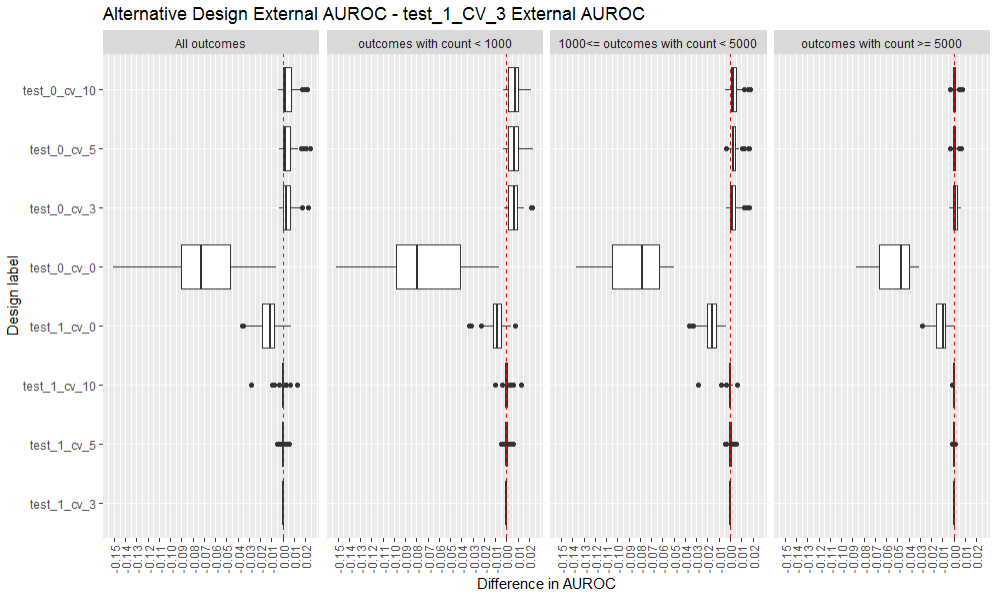


Figure 3: Box plots showing the differences in the AUROC performance when externally validated in MDCR between the design that used a test set and 3-fold CV (test_1_CV_3) vs all the other models designs across all 21 outcomes and 5 repetitions. The x-axis is the difference in AUROC and the y-axis is the design. The first column contains box plots for all outcomes, and then remaining columns group outcomes into those with a count < 1000, those with a count between 1000 and 5000 and those with a count >= 5000. These enable the effect of outcome rareness to be inspected.

BMJ Open

Investigating the impact of development and internal validation design when training prognostic models using a retrospective cohort in big US observational healthcare data

Journal:	*BMJ Open*
Manuscript ID	bmjopen-2021-050146.R1
Article Type:	Original research
Date Submitted by the Author:	18-Aug-2021
Complete List of Authors:	Reps, Jenna; Janssen Research and Development LLC; Observational Health Data Sciences and Informatics Community Ryan, Patrick; Janssen Research and Development LLC; Observational Health Data Sciences and Informatics Community Rijnbeek, P; Erasmus University Rotterdam; Observational Health Data Sciences and Informatics Community
Primary Subject Heading:	Health informatics
Secondary Subject Heading:	Research methods
Keywords:	PREVENTIVE MEDICINE, STATISTICS & RESEARCH METHODS, Health informatics < BIOTECHNOLOGY & BIOINFORMATICS

I, the Submitting Author has the right to grant and does grant on behalf of all authors of the Work (as defined in the below author licence), an exclusive licence and/or a non-exclusive licence for contributions from authors who are: i) UK Crown employees; ii) where BMJ has agreed a CC-BY licence shall apply, and/or iii) in accordance with the terms applicable for US Federal Government officers or employees acting as part of their official duties; on a worldwide, perpetual, irrevocable, royalty-free basis to BMJ Publishing Group Ltd ("BMJ") its licensees and where the relevant Journal is co-owned by BMJ to the co-owners of the Journal, to publish the Work in this journal and any other BMJ products and to exploit all rights, as set out in our licence.

The Submitting Author accepts and understands that any supply made under these terms is made by BMJ to the Submitting Author unless you are acting as an employee on behalf of your employer or a postgraduate student of an affiliated institution which is paying any applicable article publishing charge ("APC") for Open Access articles. Where the Submitting Author wishes to make the Work available on an Open Access basis (and intends to pay the relevant APC), the terms of reuse of such Open Access shall be governed by a Creative Commons licence – details of these licences and which Creative Commons licence will apply to this Work are set out in our licence referred to above.

Other than as permitted in any relevant BMJ Author's Self Archiving Policies, I confirm this Work has not been accepted for publication elsewhere, is not being considered for publication elsewhere and does not duplicate material already published. I confirm all authors consent to publication of this Work and authorise the granting of this licence.

Investigating the impact of development and internal validation design when training prognostic models using a retrospective cohort in big US observational healthcare data

Jenna M. Reps, PhD[1,2*], Patrick B. Ryan[1,2], Peter R. Rijnbeek[1,3]

[1]Observational Health Data Sciences and Informatics Community, New York, NY, USA; [2]Janssen Research and Development, Raritan, NJ, USA; [3]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

***Corresponding author - Email: jreps@its.jnj.com**

Abstract

Objective: The internal validation of prediction models aims to quantify the generalizability of a model. We aim to determine the impact, if any, that the choice of development and internal validation design has on the internal performance bias and model generalizability in big data (n ~ 500,000).

Design: Retrospective cohort

Setting: Primary and secondary care; three US claims databases

Participants: 1,200,769 patients pharmaceutically treated for their first occurrence of depression

Methods: We investigated the impact of the development/validation design across 21 real-world prediction questions. Model discrimination and calibration were assessed. We trained LASSO logistic regression models using US claims data and internally validated the models using eight different designs: 'no test/validation set', 'test/validation set' and cross validation with 3-,5- or 10-folds with and without a test set. We then externally validated each model in two new US claims databases. We estimated the internal validation bias per design by empirically comparing the differences between the estimated internal performance and external performance.

Results: The differences between the models' internal estimated performances and external performances were largest for the 'no test/validation set' design. This indicates even with large data the 'no test/validation set' design causes models to overfit. The seven alternative designs included some validation process to select the hyper-parameters and a fair testing process to estimate internal performance. These designs had similar internal performance estimates and performed similarly when externally validated in the two external databases.

Conclusions: Even with big data it is important to use some validation process to select the optimal hyper-parameters and fairly assess internal validation using a test set or cross validation.

Strengths and limitations of this study:

- We developed and externally validated 840 prediction models using 8 different development/internal validation designs across 21 prediction problems
- We focused on a target population of approximately 500,000 patients and predicted 21 different outcomes of various rareness
- We empirically investigated the impact of development/internal validation design on internal discrimination estimate bias and model generalizability in big data

Word count: 3828

Keywords:

patient-level prediction; internal validation; *sample size; best practices*

Background

Prognostic models aim to use a patient's current medical state, such as his medical history and demographics, to calculate a personalized estimate for the risk of some future medical event. If a model can make accurate predictions, then it can be used to help personalize medical decision making [1]. Big observational health care databases may provide a way to observe and follow large at-risk patient samples that could be used to develop prognostic models [2]. The initial step when using these datasets to learn a prognostic model is creating labelled data that can be used by binary classifiers. The labelled data consist of pairs of features and the outcome class for each patient in the at-risk patient sample.

Binary classification is a type of machine learning where labelled data are used to learn a model that can discriminate between two classes (e.g., healthy vs unhealthy or will develop cancer vs will be cancer free) using patient features such as age, body mass index or a medical illness (also known as attributes, predictors, or covariates). In terms of prognostic models in healthcare, a model uses current features of an at-risk patient to predict some future health state for the patient. It is hoped that a model learned using labelled data from a sample of at-risk people will generalize to any new at-risk person. Unfortunately, sometimes a model incorrectly mistakes noise in the sample of labelled data as patterns. This is known as 'overfitting' and causes a model to appear to perform extremely well in the sample of labelled data but performs much worse when applied to new data [3]. This means that the model makes incorrect predictions that could be dangerous. One way to address the issue of overfitting when developing a model is to 'hold out' some of the labelled data when learning the model and then evaluate the model on the held-out data. This process mimics evaluating the model in new data but reduces the size of the labelled data used to learn the model. Alternatively, the amount of overfitting can be quantified based on how stable the model performance is across different labelled data samples used to develop the model. This process is known as bootstrapping [4]. Using the correct internal validation design is important as it results in more reliable model performance estimates and makes it possible to fairly assess a prognostic model. Research has shown that a bootstrapped approach is most suitable in smaller datasets (<10,000 at-risk patients and <100 features) [5-6] but there is currently no research into the impact of validation design in data with a large at-risk sample (big n) and many features (big p). As healthcare datasets are growing, the a-risk samples used for model development are increasing, and the research insights found on smaller data may not extrapolate to big n and big p data. Research into the impact of development/validation design in big data is needed to ensure the most optimal models are being developed or limitations of certain designs are known.

Bootstrapping is the best approach to fairly evaluate a logistic regression model with small data due to the 'held-out' data being small and estimates being uncertain. In big n and big p data, training a model is often a slow process. Advanced machine learning methods such as deep learning can take days or weeks to train. This makes the bootstrap approach unsuitable as it requires training a model 100s of times. In addition, in big n data, the development and 'held-out' data are both large, which may overcome the small data issue of estimates being uncertain. However, as the number of features (p) increases and more complex classifiers are trained, the chance of overfitting increases, so issues may still occur in big data. Classifiers often have hyper-parameter that control the complexity. For example, regularized logistic regression models have a hyper-parameter that adds a cost to the number of features (or size of the coefficients). This makes them suitable for learning in big p data, but the optimal hyper-parameter

needs to be identified. Identifying the optimal hyper-parameters requires comparing hyper-parameter performance in some labelled data that were not used to develop the model, otherwise overfitting may bias the hyper-parameter evaluation. This means developing models in big p and big n data requires three data splits: the development data used to train the model, the validation data used to select the optimal hyper-parameter and the test data that is held out and used to fairly evaluate the model.

The bigger the data used to develop a model, the less likely the model will overfit and the bigger the 'held-out' data used to evaluate a model the more stable the performance estimates. This prompts the idea of cross-validation. Cross-validation requires splitting the labelled data into N independent subsets (N-folds) and then iterates over the subset by holding the subset out and developing the model using the combination of the N-1 other data subsets. The held-out dataset is then used to evaluate the model. This results in N performance estimates that are aggregated to provide a single estimate of performance. This provides a fair way to evaluate the model while also increasing the size of data used to develop the model. Cross-validation is often used to pick the optimal hyper-parameters. In big n and big p data there is the choice of whether to use a held-out data set (test set), whether to use a validation set or cross-validation and how many cross-validation folds to use. The common designs used for big data are displayed in Figure 1.

In this paper we compare the impact of model development design on regularized logistic regression performance in big data. We focus on data with approximately 500,000 patients, >86,000 features and investigate performance estimates across 21 prediction problems with varying outcome event count rareness. We implemented eight different development designs per prediction problem. We repeated each design multiple times with different splits (folds or test sets) to estimate how stable and unbiased the performance estimates are. We then investigate whether the choice of design impacts model performance when externally validating the models in two new databases.

Methods

We use the OHDSI PatientLevelPrediction framework [7] and R package to develop and evaluate the prediction models in this study.

Data

We developed models using a US claims database, IBM MarketScan Commercial Claims (CCAE), that contains insurance claims data for individuals enrolled in US employer-sponsored insurance health plans. The data includes adjudicated health insurance claims (e.g., inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. The patients in this database are aged under 65. The database contains records for approximately 153 million patients between Jan 2000 to Dec 2019.

Models were externally validated using:

1) IBM MarketScan Medicare Supplemental Database (MDCR), a US claims database that represents health services of retirees (aged 65 or older) in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health

plans. These data include adjudicated health insurance claims (e.g., inpatient, outpatient, and outpatient pharmacy). The database contains approximately 10 million patients from Jan 2000 to Jan 2020.
2) IBM MarketScan Multi-state Medicaid Database (MDCD), a US database containing adjudicated US health insurance claims for Medicaid enrollees from multiple states. The database includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity and Medicare eligibility. The database contains approximately 31 million patients from Jan 2006 to Jan 2020.

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

Patient and Public Involvement

No patient involved

Study Population

We extracted data for patients who are pharmaceutically treated for their first occurrence of depression to predict 21 outcomes occurring for the first time from 1 day after their depression diagnosis until 365 days after. In the development data we randomly sampled 500,000 patients from 1,964,494 treated for depression and this resulted in a range of outcome event count sizes during the 1-year follow-up. In the external validation data, we used all the data available, this corresponded to 160,956 patients in MDCR and 539,813 in MDCD.

Outcomes

We used the same 21 outcomes used by the PatientLevelPredicion framework study [7]. Table 1 lists the 21 outcomes we predicted occurring 1 day after index until 365 days after index. The number of outcome events in the development data and validation data are also reported. As we are predicting first occurrence of each outcome, we excluded patients with the outcome prior to their depression, so the study populations slightly differed per outcome (e.g., when predicting acute liver injury we exclude patients with a history of acute liver injury but when predicting ischemic stroke we exclude patients with a history of ischemic stroke).

Table 1- Outcomes predicted in this study and the logic used to define the outcome in the data.

Outcome	Phenotype	Event count in development data (N ~ 500,000)	Event count in MDCR data (N ~ 160,956)	Event count in MDCD data (N ~ 539,813)
Open angle glaucoma	A first-time condition record of Open-angle glaucoma with at least 1 condition record of Open-angle glaucoma from a provider with Ophthalmology, Optometry or	174	510	102

	Optician speciality within 1 to 365 days.			
Acute liver injury	A first-time condition record of Acute liver injury during an ER visit or inpatient visit. No Acute liver injury exclusions 1 year prior to 60 days after.	184	67	352
Ventricular arrhythmia and sudden cardiac death	A first-time condition record of Ventricular arrhythmia and sudden cardiac death during an ER visit or inpatient visit being the primary cause of the visit.	297	642	1188
Ischemic stroke	A first-time condition record of Ischemic stroke during an inpatient visit	380	1153	674
Acute myocardial infarction	A first-time condition record of Acute myocardial infarction during an ER visit or inpatient visit being the primary cause of the visit.	491	1080	1042
Gastrointestinal hemhorrage	A first-time condition record of Gastrointestinal hemhorrage during an ER visit or inpatient visit being the primary cause of the visit.	509	963	1037
Delirium	A first-time condition record of Delirium during an ER visit or inpatient visit	985	1298	1842
Seizure	A first-time condition record of Seizure during an ER visit or inpatient visit	1494	935	4314
Decreased libido	A first-time condition record of Decreased libido	1661	130	926
Alopecia	A first-time condition record of Alopecia	2577	748	2674
Hyponatremia	A first-time condition record of Hyponatremia or a first-time measurement of serum sodium between 1 and 136 millimole per litre	2628	4276	6035
Fracture	A first-time condition record of Fracture	2722	4071	4692

Vertigo	A first-time condition record of Vertigo	3046	2086	2791
Tinnitus	A first-time condition record of Tinnitus	3120	1824	3186
Hypotension	A first-time condition record of Hypotension	4170	6399	10738
Hypothyroidism	A condition record of Hypothyroidism with another condition record of Hypothyroidism within 90 days	6117	3853	6064
Suicide and suicidal ideation	A first-time condition record of Suicide and suicidal ideation or a first-time observation of Suicide and suicidal ideation	10221	993	24972
Constipation	A first-time condition record of Constipation	10672	7569	23463
Diarrhea	A first-time condition record of Diarrhea	14875	7226	24941
Nausea	A first-time condition record of Nausea	19754	7824	38344
Insomnia	A first-time condition record of Insomnia	20806	6846	32118

Candidate Predictors

We used one-hot encoding for any medical event, drug, procedure, observation or measurement recorded within 1 year prior to, or on, index (date of depression). This means we have a binary predictor per medical event/drug/procedure/observation/measurement recorded for any patient in our development study population within 1-year prior to index. For example, if a patient had a record of 'type 2 diabetes' 80 days prior to index, the value for the predictor 'type 2 diabetes 1-year prior' would be 1. If a patient never had type 2 diabetes recorded, their value for the predictor 'type 2 diabetes 1-year prior' would be 0. We also created one-hot encoded variables for any medical event, drug, procedure, observation, or measurement recorded within 30 days prior to, or on, index. In addition, we added one-hot encoded variables for age in 5-year groups (0-4,5-9,..., 95-99), index month (for seasonality), ethnicity, race and gender. Finally, the number of visits in the prior 30 days was also used as a candidate predictor. This resulted in approximately 86,000 candidate predictors. In this paper we focus on the impact of study design on internal validation estimation and therefore do not present the final developed models.

Model Development Designs

We investigate developing and internally validating LASSO logistic regression models [8] using the designs in Table 2. LASSO logistic regression is a generalized linear model that adds a penalty term to penalize the inclusion of predictors that are only weakly associated to the class label. This effectively performs feature selection during model training and is necessary due to using >86,000 candidate predictors. Due to the penalty term, only a small selection of predictors ends up being included in the final model. This makes the model less likely to overfit. The penalty amount is a hyper-parameter that needs to be determined while training the model.

We compare the estimated internal validation when developing models using:

- No test/validation set: the hyper-parameters, final model and performance are determined using all the data. This has a high risk of overestimating the performance and is included as a worst-case scenario.
- Test/validation set: the hyper-parameters are selected using the validation data, the model is fit using the training data and the performance is estimated using the test set. This is the quickest design apart from the no test/validation set.
- N-fold CV: CV on all the data is used to select the hyper-parameters and estimate the performance. Final model is fit using all the data.
- N-fold CV with test set: CV on the training data is used to select the hyper-parameters and the model is fit using all the training data. Performance is estimated using the test set.

The designs are summarized in Table 2. We investigate the impact of the number of folds (N is 3, 5 or 10) when performing CV. All designs that use CV to select the optimal hyper-parameters used the same hyper-parameter grid search. The test/train splits were done stratified by outcome, so the % of people in the test/train data with the outcome were the same.

Table 2- the different designs compared in this study.

Design	CV	Test set?	Hyper-parameter selection	Model development	Internal validation
No test/validation set	0	No	Using all data	Using all data	Using all data
Test/Validation set	0	Yes	Using 10% validation data	Using 80% training data	Using 10% test data
3-fold CV	3	No	Using 3-fold CV on all data	Using all data	Using 3-fold CV on all data
3-fold CV with test set	3	Yes	Using 3-fold CV on 80% training data	Using 80% training data	Using 20% test data
5-fold CV	5	No	Using 5-fold CV on all data	Using all data	Using 5-fold CV on all data
5-fold CV with test set	5	Yes	Using 5-fold CV on 80% training data	Using 80% training data	Using 20% test data

10-fold CV	10	No	Using 10-fold CV on all data	Using all data	Using 10-fold CV on all data
10-fold CV with test set	10	Yes	Using 10-fold CV on 80% training data	Using 80% training data	Using 20% test data

Evaluation of Models

Discrimination: The area under the receiver operating curve (AUROC) and area under the precision recall curve (AUPRC) were used to evaluate the discriminative performance (how well it ranks based on predicted risk). The AUROC is a measure that ranges between 0 and 1, with values less than 0.5 corresponding to discrimination worse than randomly guessing risk (e.g., patients who will experience the outcome are assigned a lower risk than patients who will not experience the outcome), a value of 0.5 corresponding to randomly guessing the risk and values great than 0.5 corresponding to better than random guessing. The closer the AUROC is to 1, the better the discrimination. For the AUROC estimated using N-fold CV we have N estimates of the AUROC (per fold). We calculate the 95% confidence interval (CI) using the formula mean – 1.96*standard deviation of the N estimates. For the test set AUROC we calculated the 95% CI using the standard deviation based on the Mann-Whitney statistic. The AUPRC is a measure of discrimination that is impacted by how rare the outcome is. It is the area under the curve representing the precision (probability a patient predicted as having the outcome in the future will have the outcome) as a function of recall (aka sensitivity – proportion of patient who will experience the outcome that are correctly predicted to). AUPRC also ranges between 0 to 1, with 1 representing perfect discrimination and 0 poor discrimination. However, a 'good' AUPRC value depends on the outcome proportion, and this is prediction task specific.

Calibration: To measure the calibration of the model we calculated the average E-statistic [9]. This value corresponds to the mean absolute calibration error (difference between the observed risk using a LOESS function and predicted risk). A smaller value indicates better calibration, a value of 0 means perfect calibration. The E-statistic is impacted by the outcome rareness, as a model predicting a rarer outcome will often predict lower risks and this will result in the mean error being smaller.

Model generalizability

To investigate whether some development/validation designs are more likely to cause a model to overfit (leading to optimistic internal performance estimates and making it less generalizable) we externally validated the models in two databases. The two external databases differ from the development database, so we expect some differences in model discrimination and calibration when externally validating the models. The MDCR database contains an older population and the MDCD database contains patients with a lower social economic status.

Although we expect some differences in the internal vs external performance due to data differences, very large decreases in performance when a model is applied externally may indicate that the model has

overfit. To investigate this, we calculate the difference between the internal performance compared to the external performance. A higher value for the AUROC/AUPRC discrimination metric means better discrimination, so an overfit model will have a higher internal AUROC/AUPRC then external AUROC/AUPRC. The difference, internal AUROC/AUPRC - external AUROC/AUPRC, gives an indication of whether a model has overfit to the development dataset, where a value close to zero or less than zero indicates excellent model generalizability. A lower value for the E-statistic calibration metric means better calibration, so positive internal E-statistic – external E-statistic values indicate better calibration when externally validated.

Results

The characteristics of the development and validation study populations are displayed in Table 3. The MDCR data patients were older with more comorbidities than the development data. The MDCD data patients were slightly younger and had slightly more comorbidities than the development data. The gender ratio was similar across datasets with ~70% female. The mean prior observation (number of days a patient has been active in the database prior to index) was > 1200 days (> 3 years) in all databases.

Table 3 - the characteristics of the study populations

	Development Data (N ~ 500,000)	*MDCR Data (N ~* 160,956)	*MDCD Data (N ~* 539,813*)*
Mean Age in years (sd)	*40 (15)*	*75 (7.8)*	*34 (16.6)*
Male Gender %	*31%*	*32%*	*27.1%*
Mean days prior observation (sd)	*1474 (1205)*	*1585 (1192)*	*1244 (885)*
Condition recorded in prior year (% of patients)			
Neoplastic Disease	*21.1%*	*45.7%*	*13.4%*
Pain	*60.1%*	*74.4%*	*72.8%*
Anxiety	*41.3%*	*28.6%*	*50.8%*
Respiratory tract infection	*15.9%*	*12.0%*	*22.2%*
Dementia	*0.0%*	*0.9%*	*0.1%*
Obesity	*10.5%*	*10.6%*	*17.9%*
Diabetes Mellitus	*8.9%*	*27.0%*	*13.5%*
Hypertensive disorder	*24.7%*	*69.0%*	*29.4%*
Heart disease	*9.2%*	*46.5%*	*14.0%*
Hyperlipidaemia	*23.3%*	*56.3%*	*19.8%*

Figure 2 part A displays the results of the AUROC values and 95% CI across designs for five reputations of using a test set internal validation design (in red) and using a CV internal validation or all data (in blue). The rows correspond to the number of folds used by cross-validation and the columns correspond to the

21 different outcomes. The rarest outcomes are on the left and the most common are on the right. The performance when cross-validation was not used to select hyper-parameters is the top row (no CV). In this row the 'no test/validation set' design (blue) had no validation or test set but the 'test/validation set' design (red) had a single validation set to select the hyper-parameter and a test set. Blue represents the AUC performances for designs where all the data (with or without CV) are used to estimate the internal performance, red represents the AUC performances of designs where a test set is used to estimate the internal performance and black/grey represents the external validation for each model across designs. The top row (no CV) differs from the rows 2 to 4, where we see that the 'no test/validation design' that picks the hyper-parameter and fits the model using all the same data lead to highly overfit models. The AUROC performance varied across the outcomes. In general, the external validation on MDCR (black cross) was lower than all internal validation estimates, except for three outcomes (decreased libido, alopecia and Hypothyroidism). The external validation on MDCD (grey pointer) showed the external AUROC fluctuated around the internal AUROC. The internal validation estimates using a test set vs CV appear to be similar across outcomes and the external validation performances were often equivalent across designs. The number of folds used in CV (3, 5 or 10) does not appear to impact the internal or external validation estimates, except for rare outcomes where the CIs are wider. Similar trends were observed when considering the AUPRC and E-statistic, see Figure 2 part B and Figure 2 part C.

To investigate whether some development/validation designs are more likely to lead to optimistic internal discriminative estimates we calculated the difference between the internal validation performance and the external validation performance in MDCD and MDCR for each model. Figure 3 shows box plots for the difference between the internal performance and the external performance on the x-axis with the y-axis representing the design used to develop/validate each model. The red box plots are the differences when externally validated in MDCD and the blue box plots are differences when externally validated in MDCR. The AUROC, AUPRC and E-statistic performance metrics differences are displayed. The results show that the 'no test/validation design' resulted in optimistic AUROC and AUPRC, as the differences were large in both databases. The design also resulted in worse external calibration. The other designs had similar difference distributions in Figure 3 and similar performances in Figure 2.

To see whether these results are consistent across different outcome counts, we also include the difference distributions broken up by prediction tasks with an outcome count less than 1000, outcome count between 1000 and 5000 and outcome count of 5000 or more, see Supplement Figures s1-s3. The difference distributions were similar across all three metrics. Figure 2 part A shows that when the outcome count is < 1500, the AUROC performance fluctuated per replication for all designs except the overfit 'no test/validation set' design.

Discussion

In small data it has been shown that the design used to development and internal validated a model impacts the internal performance estimate bias. In this study using big n (500,000) and big p (>86,000) data to develop LASSO logistic regression models we show that the impact of design has negligible impact if some fair validation process is implemented to select the optimal hyper-parameter and some fair process is implemented to estimate the internal performances. The only design in this study that resulted in highly biased internal performance estimates was the 'no test/validation' design that leads to

overfit models even with big data. The estimated performance of any prognostic model that is developed using the 'no test/validation' design cannot be trusted, and this design should be avoided.

Interestingly in this study the number of folds used by CV appeared to have negligible impact on the model's internal and external performance in big data. This is a useful result, as increasing the number of folds makes the model development more complex and could slow down model development.

We sampled 500,000 target patients from the development database to reduce the lower value of outcome count range across the 21 outcomes. This enabled us to gain insight into the impact of low outcome count on the internal performance estimate per design. The number of outcomes has been shown to impact model performance [10]. We can see from Figure 2 that the split used to create the data used for selecting the hyper-parameter and evaluating the model impacted the internal AUROC estimates when the outcome count was <1500 as the values varied across replication. This suggests that even in big data (n =500,000) and using an appropriate design, if the outcome is rare (< 0.3%) the internal validation will have some error. The designs that used CV rather than a test set to estimate internal performance were more stable when the outcome was less common. This makes sense as holding out data for a test set reduces the amount of data used to develop the model and this will have an impact on performance if the outcome count is low.

The AUROC is not impacted by outcome rareness, so the difference in internal and external AUROC represents the difference in discriminative ability of the model in the development data and the external databases. The AUPRC and E-statistic are impacted by the outcome rareness, so differences between the internal and external performances for these metrics were impacted by differences in the outcome rate in the development data and external data. This explains why the AUPRC was often greater when models were applied to the external data.

The main strength of this study is that we were able to investigate the impact of development/validation design across a large number of outcomes. In total we investigated 8 designs no test/validation set, test/validation set and 3-fold/5-fold/10-fold CV with/without a test set, 21 outcomes and 5 repetitions, resulting in the development of 840 models (8x21x5). In addition, we externally validated each of these models in two different databases. The percentage of the study population experiencing each outcome ranged between ~0.04% to ~4%, enabling us to investigate the impact of development/validation design in big data when the outcome count was small and large.

Limitations of this study include only investigating models developed in one US claims data and in future work it would be useful to repeat this study using more datasets to see whether the results hold. Similarly, we only used one target population, patients initially treated for depression, and future work should investigate whether the results hold across different study populations and outcomes. Finally, we have only investigated the impact of the model development design when developing a LASSO logistic regression. Our results may not generalize to all binary classifiers.

Conclusion

Our study is the first to investigate the impact of model development/validation design on the accuracy of the internal discrimination/calibration estimate and external validation performance when using big data (n=500,000). We compared designs that use i) all the data to develop and validate a model (no test/validation set), ii) a train/test/validation set (test/validation set), iii) CV with a test set and iv) CV only to estimate the internal discriminative performance across 21 prediction problems. The results showed that the 'no test/validation set' design leads to overfitted models that have unrealistically high internal discrimination estimates but the other designs were able to limit overfitting equivalently. These results show that even in big data using a poor design to develop LASSO logistic regression models can impact the accuracy of the internal validation and compromise model generalizability. A useful design requires: 1) a fair process to pick any hyper-parameters (e.g., a validation set or CV) and 2) a fair process to evaluate the model internally (e.g., a test set or CV).

Funding statement
This work was supported by the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

Ethics approval statement
The use of IMB CCAE, IBM MDCR and IBM MDCD was reviewed by the New England Institutional Review Board (IRB) and were determined to be exempt from broad IRB approval. Only de-identified data were used, and informed consent was not applicable.

Competing interests statement
Dr. Reps and Dr. Ryan report and are employees of Janssen Research and Development and are shareholders of Johnson & Johnson

Dr. Rijnbeek reports grants from Innovative Medicines Initiative, grants from Janssen Research and development, during the conduct of the study.

Author's contribution
JMR and PBR contributed to the conception and design of the work, JMR implemented the analysis. JMR, PBR and PRR contributed to the interpretation of data for the work and in drafting, revising and approving the final version.

Data sharing statement

The IBM CCAE, IBM MDCD and IBM MDCR data that support the findings of this study are available from IBM MarketScan Research Databases (contact at: http://www.ibm.com/us-en/marketplace/marketscan-researchdatabases) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

References

1. Steyerberg, E.W., Moons, K.G., van der Windt, D.A., Hayden, J.A., Perel, P., Schroter, S., Riley, R.D., Hemingway, H., Altman, D.G. and PROGRESS Group, 2013. Prognosis Research Strategy (PROGRESS) 3: prognostic model research. *PLoS medicine*, *10*(2), p.e1001381.

2. Goldstein, B.A., Navar, A.M., Pencina, M.J. and Ioannidis, J., 2017. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, *24*(1), pp.198-208.

3. Ying, X., 2019, February. An overview of overfitting and its solutions. In *Journal of Physics: Conference Series* (Vol. 1168, No. 2, p. 022022). IOP Publishing.

4. Efron, B. and Tibshirani, R.J., 1994. *An introduction to the bootstrap*. CRC press.

5. Steyerberg, E.W., Harrell Jr, F.E., Borsboom, G.J., Eijkemans, M.J.C., Vergouwe, Y. and Habbema, J.D.F., 2001. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *Journal of clinical epidemiology*, *54*(8), pp.774-781.

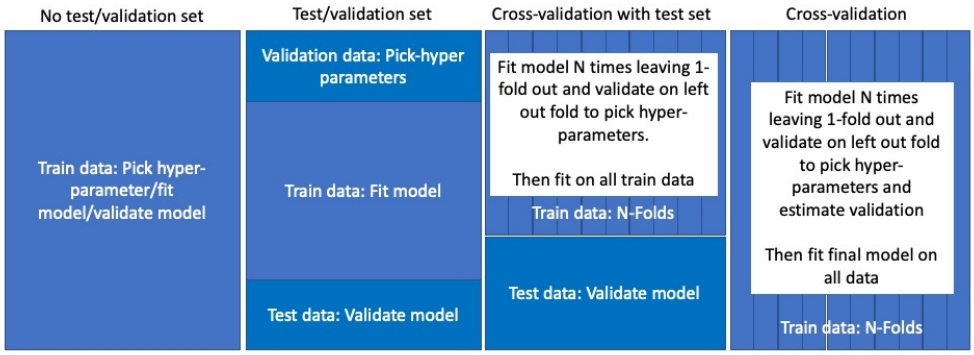
6. Steyerberg, E.W. and Harrell, F.E., 2016. Prediction models need appropriate internal, internal–external, and external validation. *Journal of clinical epidemiology*, *69*, pp.245-247.

7. Reps JM, 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):969-75.

8. Suchard MA, Simpson SE, Zorych I, et al. Massive parallelization of serial inference algorithms for complex generalized linear models. *ACM Transactions on Modeling and Computer Simulation*, 2013;23(1):10-32.

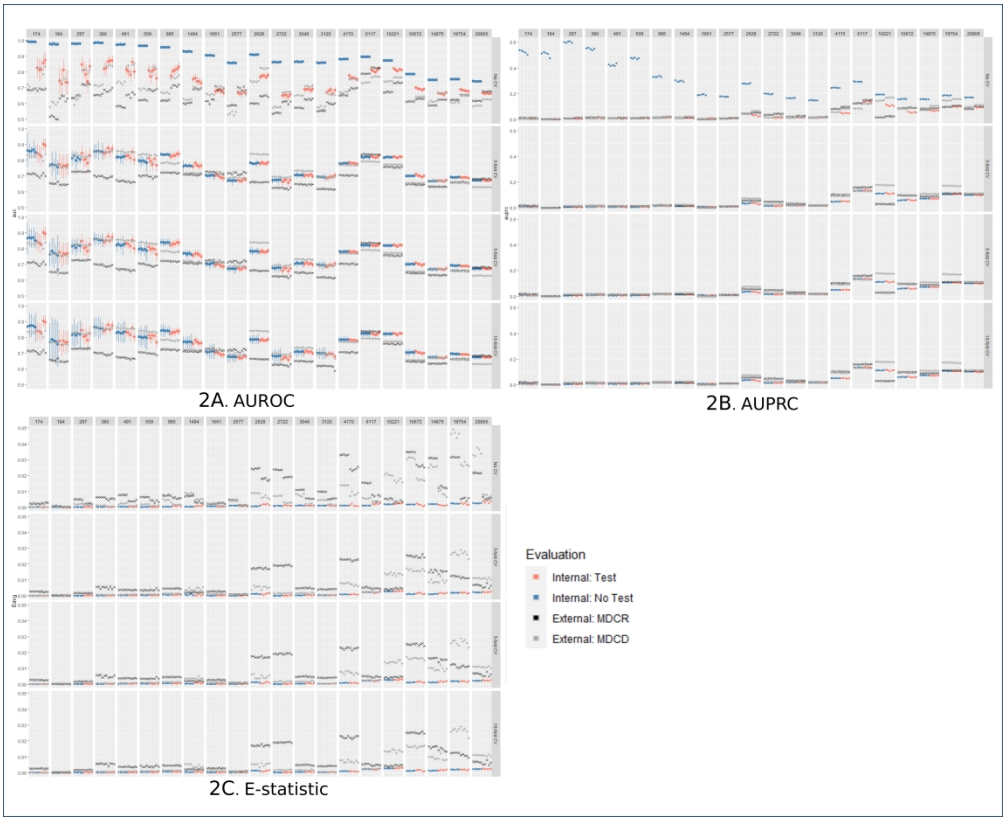
9. Harrell FE Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York, NY: Springer-Verlag; 2001.

10. John, L.H., Kors, J.A., Reps, J.M., Ryan, P.B. and Rijnbeek, P.R., 2020. How little data do we need for patient-level prediction?. *arXiv preprint arXiv:2008.07361*.



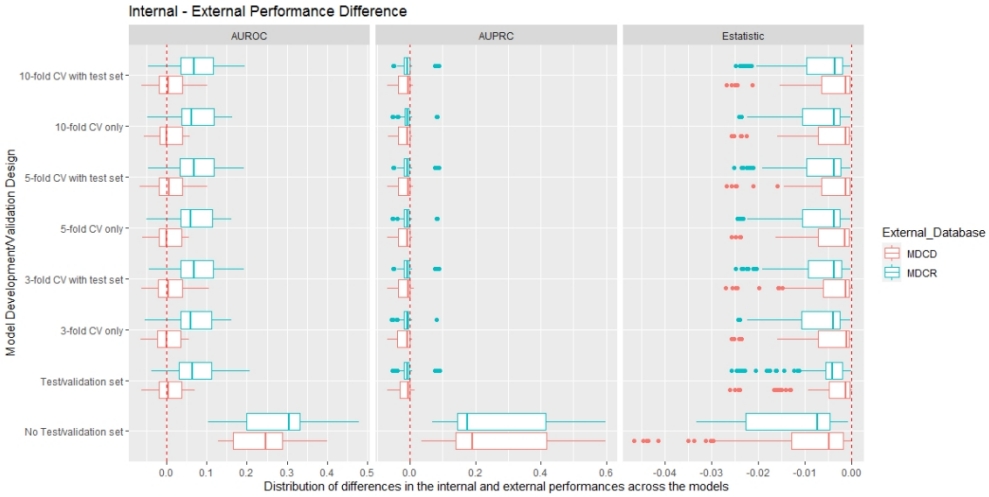
Possible development and internal validation design strategies for big data. The options include whether to use a test set (hold out some data from development that is used to fairly assess performance) and whether to use cross validation (where the data are partitioned, and each partition is iteratively held out while the rest of the data are used to develop the model).

338x190mm (72 x 72 DPI)



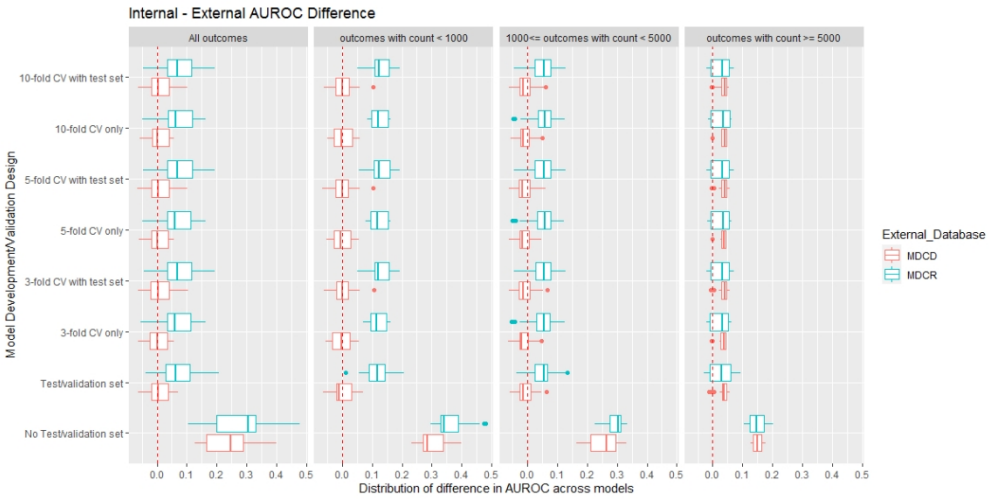
The AUROC/AUPRC/E-statistic performance estimates for five repetitions per design per prediction task. The columns represent the prediction task, with the number representing the number of patients with the outcome during the time-at-risk. For example, the first column corresponds to a prediction task where 174 patients had the outcome, whereas the last column corresponds to a prediction task where 20806 patients had the outcome. The rows correspond to whether CV was used by the design (top row does not use CV) or the number of folds (3, 5 or 10). The internal validation performances of the designs that used a test set are colored in red, and those not using a test set are blue. The external validation performances for a model are the grey markers (MDCD) and black crosses (MDCR) that have the same x-coordinate and fall within the same row/column.

819x666mm (157 x 157 DPI)



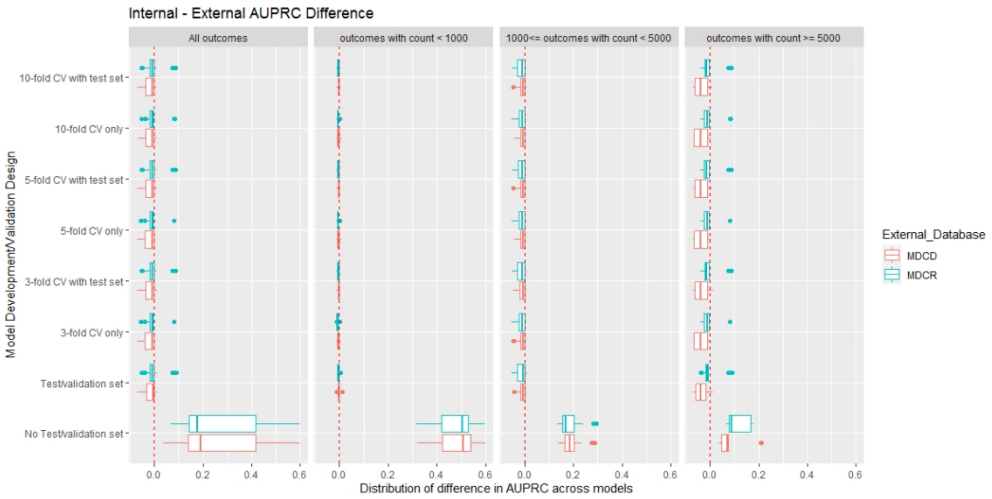
Box plots showing the internal performance estimate minus the external performance estimate per design and external database. The left side shows the AUROC differences, the center shows the AUPRC differences, and the right side shows the E-statistic differences. For the AUROC, values near 0 indicate that the internal validation AUROC estimates were accurate as the external validation AUROCs were similar. For AUPRC and AUPRC values less than 0 indicate that the performance was better externally, values greater than 0 indicate the performance is worse externally. For the E-statistic, values less than 0 indicate worse calibration when the models were externally validated.

802x401mm (38 x 38 DPI)



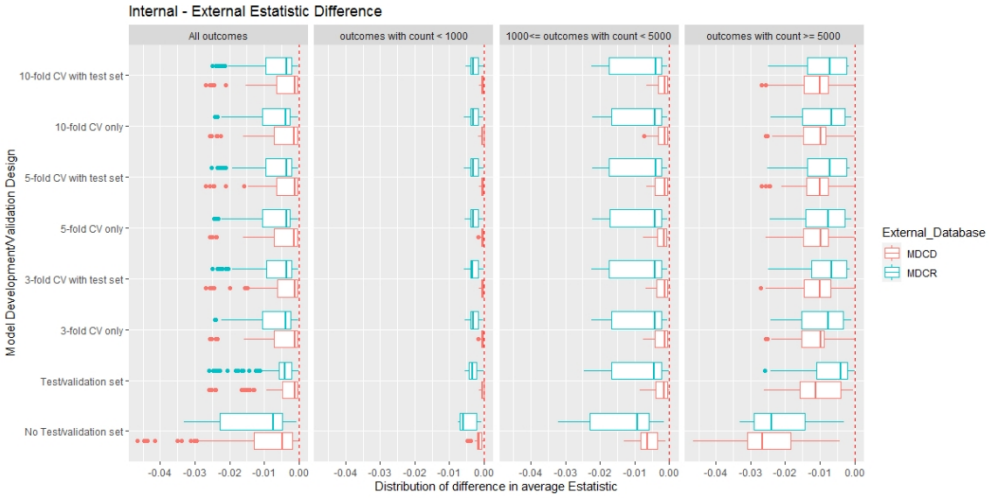
Box plots showing the Internal AUROC estimate minus the external AUROC estimate per design. Values near 0 indicate that the internal validation AUROC estimates were accurate as the external validation AUROCs were similar. The first column contains box plots for all outcomes, and then remaining columns group outcomes into those with a count < 1000, those with a count between 1000 and 5000 and those with a count >= 5000. These enable the effect of outcome rareness to be inspected.

802x401mm (38 x 38 DPI)



Box plots showing the internal AUPRC estimate minus the external AUPRC estimate per design. Values near 0 indicate that the internal validation AUPRC estimates were accurate as the external validation AUPRCs were similar. AUPRC depends on the outcome percentage in the target population, so differences could occur due to different outcome rareness between databases. The first column contains box plots for all outcomes, and then remaining columns group outcomes into those with a count < 1000, those with a count between 1000 and 5000 and those with a count >= 5000. These enable the effect of outcome rareness to be inspected.

802x401mm (38 x 38 DPI)



Box plots showing the difference between the internal E-statistic estimate and the external E-statistic estimate per design. The more negative this difference, the worse the calibration was when externally validated. E-statistic depends on the outcome percentage in the target population, so differences in calibration could occur due to different outcome rareness between databases. The first column contains box plots for all outcomes, and then remaining columns group outcomes into those with a count < 1000, those with a count between 1000 and 5000 and those with a count >= 5000. These enable the effect of outcome rareness to be inspected.

802x401mm (38 x 38 DPI)