Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets

Toshihiko Takada , Steven Nijman , Spiros Denaxas , Kym I.E. Snell , Alicia Uijl , Tri-Long Nguyen , Folkert W. Asselbergs , Thomas P.A. Debray

PII:	S0895-4356(21)00107-4
DOI:	https://doi.org/10.1016/j.jclinepi.2021.03.025
Reference:	JCE 10468

To appear in:	*Journal of Clinical Epidemiology*

Accepted date:	29 March 2021

Internal-external cross-validation helped to evaluate the generalizability of prediction models in large clustered datasets

Toshihiko Takada[a], Steven Nijman[a], Spiros Denaxas[b,c,d,e], Kym I.E. Snell[f], Alicia Uijl[a,g,h], Tri-Long Nguyen[a,i], Folkert W. Asselbergs[b,h,j], Thomas P.A. Debray[a,b,*]

[a]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht University, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands

[b]Health Data Research UK and Institute of Health Informatics, University College London, Gibbs Building, 215 Euston Road, London, NW1 2BE, United Kingdom

[c]The Alan Turing Institute, British Library, 96 Euston Road, London, NW1 2DB, United Kingdom

[d]The National Institute for Health Research University College London Hospitals Biomedical Research Centre, University College London, Suite A, 1st floor, Maple House, 149 Tottenham Court Road, London, W1T 7DN, United Kingdom

[e]British Heart Foundation Research Accelerator, University College London, Gower Street, London, WC1E 6BT, United Kingdom

[f]Centre for Prognosis Research, School of Medicine, Keele University, Keele, Staffordshire, ST5 5BG, United Kingdom

[g]Division of Cardiology, Department of Medicine, Karolinska Institute, 171 77 Stockholm, Sweden

[h]Department of Cardiology, Division Heart & Lungs, University Medical Center Utrecht, Utrecht University, Heidelberglaan 100, PO Box 85500, 3508GA, Utrecht, The Netherlands

[i]Section of Epidemiology, Department of Public Health, University of Copenhagen, CSS, Øster Farimagsgade 5, DK-1353 Copenhagen K, Denmark

[j]Institute of Cardiovascular Science, Faculty of Population Health Sciences, University College London, Gower Street, London, WC1E 6BT, United Kingdom

Declaration of interests

TT was supported by the Uehara Memorial Foundation between August 2018 and July 2019. SN is supported by a Public-Private Study grant of the Dutch Heart foundation (2018B006). KIES is supported by a National Institute for Health Research School for Primary Care Research (NIHR SPCR) Launching Fellowship. The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. FWA is supported by UCL Hospitals NIHR Biomedical Research Centre. TPAD is supported by the Netherlands Organisation for Health Research and Development (91617050) and the European Union’s Horizon 2020 Research and Innovation Programme under ReCoDID Grant Agreement No. 825746. All other authors declared no competing interests.

*** Corresponding author:**

Thomas P.A. Debray
Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht,
Utrecht University, Universiteitsweg 100, 3584 CG, Utrecht, The Netherlands
T.Debray@umcutrecht.nl
+31 88 75 680 26

Abstract

Objective

To illustrate how to evaluate the need of complex strategies for developing generalizable prediction models in large clustered datasets.

Study Design and Setting

We developed eight Cox regression models to estimate the risk of heart failure using a large population-level dataset. These models differed in the number of predictors, the functional form of the predictor effects (non-linear effects and interaction) and the estimation method (maximum likelihood and penalization). Internal-external cross-validation was used to evaluate the models' generalizability across the included general practices.

Results

Among 871,687 individuals from 225 general practices, 43,987 (5.5%) developed heart failure during a median follow-up time of 5.8 years. For discrimination, the simplest prediction model yielded a good concordance statistic, which was not much improved by adopting complex strategies. Between-practice heterogeneity in discrimination was similar in all models. For calibration, the simplest model performed satisfactorily. Although accounting for non-linear effects and interaction slightly improved the calibration slope, it also led to more heterogeneity in the observed/expected ratio. Similar results were found in a second case study involving patients with stroke.

Conclusion

In large clustered datasets, prediction model studies may adopt internal-external cross-validation to evaluate the generalizability of competing models, and to identify promising modelling strategies.

Keywords

Prediction model; Calibration; Discrimination; Validation; Heterogeneity; Model comparison

Running title: Internal-external cross-validation in large clustered datasets

Word count: 2993

What is new?

Key findings

Flexible modelling strategies did not improve prediction model performance across different settings and populations.

Although the inclusion of additional predictors marginally improved the model's discriminative performance, it also increased between-practice heterogeneity (thereby impairing model generalizability).

What this adds to what was known

In contrast to traditional internal validation methods, internal-external cross-validation (IECV) can quantify the generalizability of a prediction model across different settings and populations.

What is the implication and what should change now?

When developing prediction models using large clustered datasets, both their internal and external validity should be studied.

IECV can be used to compare the practical benefits of different modelling strategies, and to simplify model complexity.

1. Introduction

In medicine, there are an increasing number of clinical prediction models [1]. These models aim to predict a risk of having a certain condition or experiencing a health event in the future. Prediction models are often developed using a single and small dataset. This leads to prediction models that are more prone to overfitting with the dataset used for its development, which leads to poor accuracy and less generalizability of risk predictions when the model is validated or used in new individuals.

For this reason, there has been a growing interest in prediction model studies using large datasets from electronic health records (EHRs), multi-center studies or individual participant data [2–5]. An advantage of such large datasets is that parameters of the prediction model can accurately be estimated, thereby facilitating the development of complex models with many predictors, interaction terms and/or non-linear effects. Furthermore, a common feature of these large datasets is that individuals are often clustered within hospitals, primary care practices, or even within countries. Clusters may differ with respect to included participants, variable definitions and measurement methods, all of which may affect the generalizability of developed prediction models. The presence of clustering, however, also offers an important opportunity, as the performance of a prediction model can be examined on multiple occasions and thus be used to explore its generalizability across different settings and populations. Recently, various strategies for such analyses using large clustered data have been proposed [2, 5–8].

The aim of this study was to illustrate how advanced methods can be used to evaluate the need of complex strategies for developing generalizable clinical prediction models in large clustered datasets.

2. Methods

For illustration purpose, we used two case studies.

2.1. Case study 1

We compared various modelling strategies using an example of a prediction model for the incidence of heart failure (HF). In the field of cardiovascular diseases (CVD), HF is one of the most relevant outcomes due to its high morbidity and mortality [9–12].

2.1.1. Source of the data

We used an existing large population-level dataset which links three sources of EHRs in England: primary care records from the Clinical Practice Research Datalink (CPRD), secondary care diagnoses and procedures recorded during admissions in Hospital Episodes Statistics (HES), and the cause-specific death registration information sourced from the Office for National Statistics (ONS) registry. This study was carried out as part of the CALIBER © resource (https://www.ucl.ac.uk/health-informatics/caliber and https://www.caliberresearch.org/) [13, 14]. CALIBER, led from the UCL Institute of Health Informatics, is a research resource providing validated EHR phenotyping algorithms and tools for national structured data sources. Data were recorded in five controlled clinical terminologies: Read version 2 (CPRD diagnoses), International classification of diseases (ICD)-9 and ICD-10 (HES diagnoses, ONS causes of death), the Office of Population Censuses and Surveys (OPCS)-4 (HES procedures) and British National Formulary (BNF) (CPRD medication prescriptions). The study was approved by the MHRA (UK) Independent Scientific Advisory Committee (14_246RMnA2), under Section 251 (NHS Social Care Act 2006).

2.1.2. Population

The construction of this cohort has been described by Uijl et al [15]. Briefly, we selected all individuals that were 55 years or older between 1st January 2000 and 25th March 2010, and had at least one year of follow-up by a general practitioner, in a practice that had at least one year of up-to-standard data recording in CPRD. The last date of the follow-up between the period above was considered cohort entry date (index date). Individuals with a history of HF before their index date were excluded. The study flow diagram is shown in Appendix A.

2.1.3. Predictors

We identified predictors that are commonly measured in CPRD or HES, and commonly used for prediction of HF [15, 16]: Age, sex, current smoking, ethnicity (CE, Caucasian ethnicity), index of multiple deprivation (IMD), body mass index (BMI), creatinine level (CL), and total cholesterol (TC). IMD is a measure of multiple deprivation at the small area level, consisting of seven domains [17]. Within this set, we selected those predictors which were least affected by missing data. The closest measurement to index date between three years before and one year after the index date was used. Detailed information about the definition of each predictor is available on the CALIBER website [18].

2.1.4. Outcomes

The primary outcome was incidence of HF, based on the first record of HF from CPRD or HES after the index date. In CPRD, HF was defined by a diagnosis of HF or chronic left ventricular dysfunction on echocardiogram with READ codes. In HES, it was defined by a diagnosis of HF during a hospitalization using all positions of ICD-10. If no diagnosis of HF was made, censoring was defined as the first event among the following: death, de-registration from a practice, last practice data collection, or at the study end date.

2.1.5. Statistical analysis

2.1.5.1. Multilevel imputation

Multiple multilevel imputation which accounts for potential heterogeneity between the included clusters is recommended in the recent methodological guidelines [19], however, due to limited hardware processing capacity, we applied single multilevel imputation. The detail of the imputation process is described in Appendix B.

2.1.5.2. Derivation and validation of prediction models

We considered eight modelling strategies to predict the risk of developing HF using Cox regression. These models differed with respect to the number of predictors, the functional form of the predictor effects and the method of estimation. Each model and their estimation method are summarized in Table 1.

Table 1. Description of the eight prediction models

Model	Included predictor variables								2-way IT	# RC	Estimation method
	Age	Male Sex	Smoking	CE	IMD	BMI	CL	TC			
1	L	L	L	L	-	-	-	-	no	4	Cox regression
2	RCS	L	L	L	-	-	-	-	yes	14	Cox regression
3	L	L	L	L	-	-	-	-	no	4	Ridge penalized Cox
4	RCS	L	L	L	-	-	-	-	yes	14	Ridge penalized Cox
5	L	L	L	L	L	L	L	L	no	8	Cox regression
6	RCS	L	L	L	RCS	RCS	RCS	RCS	yes	66	Cox regression
7	L	L	L	L	L	L	L	L	no	8	Ridge penalized Cox
8	RCS	L	L	L	RCS	RCS	RCS	RCS	yes	66	Ridge penalized Cox

Abbreviations: IT, interaction terms. #RC, the total number of regression coefficients. CE, Caucasian ethnicity. IMD, index of multiple deprivation. BMI, body mass index. CL, creatinine level. TC, total cholesterol. L, Linear effects. RCS, restricted cubic splines.

Models 1, 3, 5 and 7 included all predictor variables as linear effects. Models 2, 4, 6 and 8 used RCS with three knots for all continuous predictor variables, and interaction terms between all possible combinations of two variables. Model 3, 4, 7, and 8 were estimated using a ridge penalty. For all models, the total number of regression coefficients is displayed.

Model 1 included four predictors (age, sex, current smoking, and CE) as linear effects. Model 2 was an extension of Model 1 that included non-linear effect for age and for all possible two-way interactions between the four predictors. Model 3 and 4 included the same predictors as Model 1 and 2, respectively, but were estimated using a ridge penalty. Model 5 was an extension of Model 1 that also included IMD, BMI, CL and TC as linear effects. Model 6 – 8 were extended from Model 5 as similar to Model 2 – 4 from Model 1. In models with a ridge penalty (Model 3, 4, 7 and 8), all regression coefficients were shrunk towards zero by penalizing the partial log-likelihood for the magnitude of the squared coefficients (L2-norm) [20]. This strategy has been recommended to avoid overfitting, and to improve prediction model performance, particularly when it is applied in new population. We used the degree of penalty (lambda) which minimized the mean square error in ten-fold cross validation. The proportional hazards assumption of all models was checked using the Schoenfeld residuals.

We performed internal-external cross-validation (IECV) to compare the performance of the aforementioned eight prediction models at multiple occasions [2, 6]. In contrast to traditional internal validation methods (e.g., bootstrapping, cross-validation) which evaluate the model's performance in new individuals from the same population (i.e., reproducibility), IECV assesses model performance in new individuals from different but related practices as compared to the original development sample. These practices (i.e., taken as cluster) may differ with respect to case-mix, variable definitions and measurement methods, and thus allow to investigate the model's generalizability [21]. Using IECV, the data from all but one practice are used for estimating the prediction model, after which its performance is evaluated in the remaining practice. The procedure is repeated by rotating the omitted practice, resulting in multiple estimates of prediction model performance. For each prediction model, we assessed the model's discrimination performance using Harrell's concordance (c-) statistic. For calibration, we constructed calibration plots in the overall population. We also estimated the calibration slope and the ratio of observed versus expected events (O:E ratio) at five years of follow-up [22]. Interpretation of each performance measure is described in Appendix C.

The performance measures resulting from IECV were pooled using random-effect meta-analysis [2, 23, 24]. This approach not only accounts for the precision of practice-specific performance estimates, but also quantifies the between-practice variability (heterogeneity) of model performance. Heterogeneity is quantified by the between-practice standard deviation of model performance (τ) [7]. Meta-analysis results were reported as point estimates with 95% confidence intervals (CI) and 95% prediction intervals (PI). The CI indicates the precision of the model's average performance across all practices. Conversely, the PI accounts for heterogeneity between practices and therefore indicates what performance can be expected when the model is applied within a specific practice.

2.2. Case study 2

In this case study, we used patient-level data from a large international, multi-center, randomized controlled trial [25]. Because the missingness proportion was very low (6.0%), we performed a complete case analysis. Eight modelling strategies using ridge penalized Cox regression model were considered to predict the risk of mortality from CVD in patients with acute ischemic stroke. These models differed with respect to the number of predictors, the functional form of the predictor effects (non-linear effects and/or interaction terms). We illustrated the advantage of IECV by comparing it with bootstrap internal validation. More detailed information is available in Appendix D.

All analyses were performed using R version 3.6.1.

3. Results

3.1. Case study 1

The cohort included 871,687 individuals from 225 general practices. Among these, 43,987 (5.5%) developed HF during a median follow-up time of 5.8 years (interquartile range [IQR] 2.7 – 9.9), with a median time-to-event of 3.7 years (IQR 1.8 – 6.4). Baseline characteristics are shown in Table 2.

Table 2. Baseline characteristics of the cohort

Predictor variable	Individuals with incident HF	Individuals without HF	Proportion of missing
Total number of patients	43,987	823,700	
Age, years, median (IQR)	75.5 (68.5 – 81.5)	60.6 (55.0 - 70.5)	0.0%

Male sex, n (%)	22,618 (51.4)	442,409 (53.7)	0.0%
Caucasian ethnicity, n (%)	42,065 (95.6)	754,756 (91.6)	39.2%
Current Smoking, n (%)	10,843 (24.7)	190,851 (23.2)	66.2%
IMD, median (IQR)	16.2 (9.4 - 27.1)	13.7 (8.3 - 23.4)	0.3%
BMI, kg/m2, median (IQR)	27.4 (23.9 - 31.0)	26.9 (23.6 - 30.4)	60.2%
Creatinine, µmol/L, median (IQR)	102.4 (85.0 - 122.4)	88.7 (73.1 - 105.6)	66.5%
Total cholesterol, mmol/L, median (IQR)	5.3 (4.6 - 6.1)	5.5 (4.8 - 6.3)	72.3%

Abbreviations: HF, heart failure. IQR, interquartile range. IMD, index of multiple deprivation. BMI, body mass index.

The number of patients with HF in each general practice was a median of 197 (IQR 128 – 282, range 3 – 622). We explored heterogeneity of case-mix across the included general practices by comparing their distribution of predicted risk according to Model 5. Results in Appendix E indicate that the standard deviation (SD) of the linear predictor (LP) in each general practice ranged between 1.09 and 1.41, and that the mean LP in each general practice ranged between -0.51 and 0.61.

The estimated regression coefficients of the eight prediction models, as obtained from the entire dataset, are presented in Appendix F. These results indicate that all included predictors were significantly associated with HF, and that interactions were present between various predictors. The performance of the estimated models, as evaluated using IECV, is summarized in Table 3.

Table 3. Meta-analysis results of the model performance

Model	# RC	Model Estimation	Summary Estimate	95% CI	95% PI	SE*	τ^*
Discrimination performance (c-statistic)							

1	4	Cox regression	0.792	0.788	0.796	0.741	0.835	0.012	0.145
2	14	Cox regression	0.793	0.789	0.797	0.742	0.836	0.012	0.144
3	4	Ridge penalized Cox	0.793	0.789	0.796	0.742	0.835	0.012	0.144
4	14	Ridge penalized Cox	0.793	0.789	0.796	0.742	0.836	0.012	0.144
5	8	Cox regression	0.808	0.804	0.812	0.756	0.852	0.012	0.156
6	66	Cox regression	0.806	0.802	0.810	0.744	0.856	0.014	0.180
7	8	Ridge penalized Cox	0.808	0.804	0.812	0.757	0.851	0.012	0.153
8	66	Ridge penalized Cox	0.809	0.805	0.813	0.754	0.854	0.013	0.163
Calibration performance (O:E ratio at 5 years)									
1	4	Cox regression	0.957	0.926	0.990	0.598	1.532	0.017	0.239
2	14	Cox regression	0.963	0.926	1.001	0.557	1.665	0.020	0.279
3	4	Ridge penalized Cox	0.959	0.928	0.991	0.609	1.511	0.017	0.231
4	14	Ridge penalized Cox	0.958	0.927	0.990	0.609	1.508	0.017	0.231
5	8	Cox regression	0.950	0.922	0.977	0.640	1.408	0.015	0.200
6	66	Cox regression	0.935	0.903	0.969	0.572	1.530	0.018	0.251
7	8	Ridge penalized Cox	0.947	0.921	0.974	0.648	1.385	0.014	0.193
8	66	Ridge penalized Cox	0.954	0.928	0.981	0.655	1.389	0.014	0.191
Calibration performance (calibration slope)									
1	4	Cox regression	1.021	1.005	1.036	0.835	1.206	0.008	0.094
2	14	Cox regression	1.010	0.992	1.028	0.789	1.231	0.009	0.112

3	4	Ridge penalized Cox	1.126	1.108	1.143	0.923	1.328	0.009	0.103
4	14	Ridge penalized Cox	1.088	1.071	1.105	0.888	1.287	0.009	0.101
5	8	Cox regression	1.023	1.007	1.039	0.833	1.214	0.008	0.097
6	66	Cox regression	0.992	0.975	1.008	0.792	1.191	0.008	0.101
7	8	Ridge penalized Cox	1.138	1.120	1.156	0.917	1.358	0.009	0.112
8	66	Ridge penalized Cox	1.077	1.061	1.092	0.892	1.261	0.008	0.094

Abbreviations: #RC, the total number of estimated regression coefficients. CI, confidence interval. PI, prediction interval. SE, standard error.

For all models, summary estimates were obtained using random effects meta-analysis. The SE and between-study heterogeneity (τ) are given on the scale of the meta-analysis (that is, the logit of c-statistic, the log of the O:E ratio and identity for the calibration slope).

3.1.1. Discrimination performance

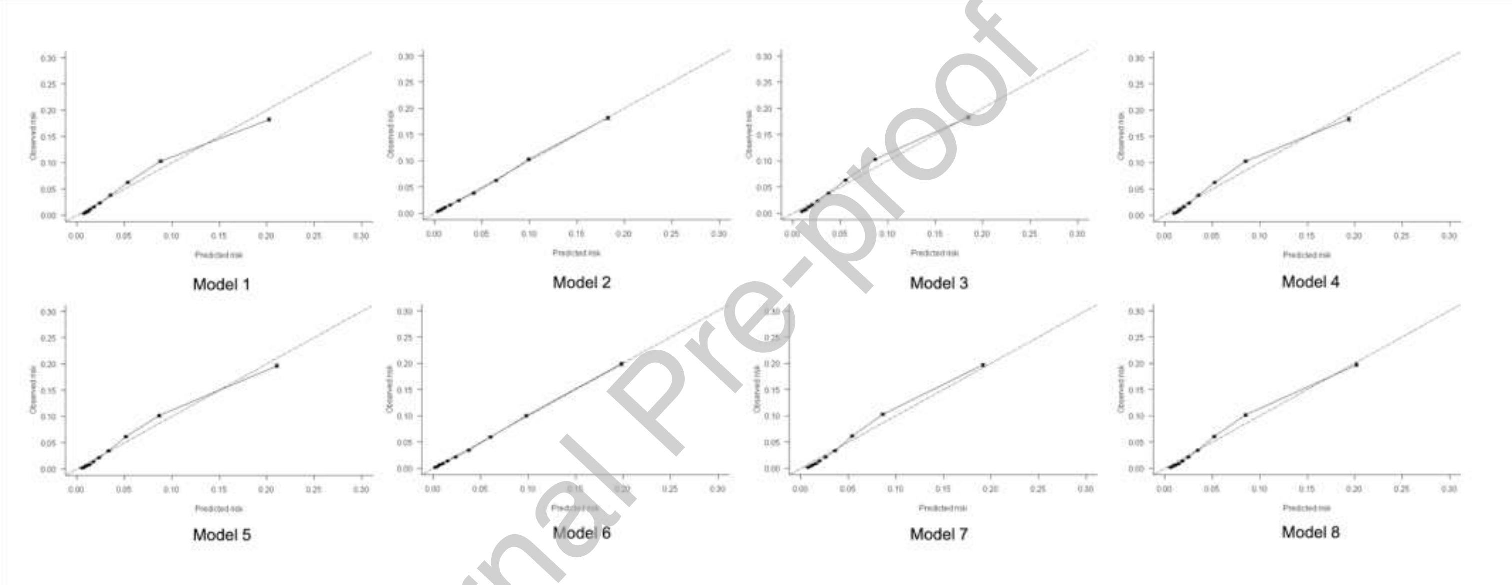
The c-statistic across the general practices is shown in Appendix G. All models showed similar discrimination, although models that included more predictors yielded somewhat larger values for the c-statistic (0.79 in Model 1 – 4 vs. 0.81 in Model 5 – 8). For all models, there was notable between-practice heterogeneity in discrimination performance. For instance, the 95% PI for a Cox regression model including eight predictors as main effects (model 5) ranged from 0.756 to 0.852. Estimates for the between-study standard deviation (τ) were similar for all models, but slightly larger for prediction models that included eight predictors and allowed for non-linear effects and interactions.

3.1.2. Calibration performance

3.1.2.1. Calibration plot

Calibration plots in Figure 1 indicate that predicted and observed risks were almost in perfect agreement for the unpenalized Cox regression model that included non-linear effects and interactions between predictors (Model 2 and 6).

Figure 1. Calibration plots of the eight prediction models



Model 1

Model 2

Model 3

Model 4

Model 5

Model 6

Model 7

Model 8

Predicted and observed risks are almost in perfect agreement for the unpenalized Cox regression models that included non-linear effects and interactions between predictors (Model 2 and 6). Some under-prediction for risk estimates around 10% is observed in the remaining models.

3.1.2.2. O:E ratio

The O:E ratio across the included general practices is shown in Appendix H. All models yielded summary O:E ratios at 5 years below one, especially those models that included eight predictors (Model 5 – 8). In addition, PIs indicate that all prediction models may substantially over- or under-predict the risk of HF when applied to individual patients from a new practice.

3.1.2.3. Calibration slope

Calibration slope across the included general practices is shown in Appendix I. Unpenalized prediction models yielded pooled calibration slopes most close to one (Model 1, 2, 5, and 6). Prediction models that adopted a ridge penalty yielded calibration slopes that were slightly larger than one, indicating that predicted risks did not vary enough and thus that too much shrinkage may have been applied in the development sample. For all models, the calibration slope was prone to a limited amount of between-practice heterogeneity. For instance, the prediction model that included eight predictors as main effects (model 5) yielded a 95% PI from 0.833 to 1.214. Estimates of between-study variance of the calibration slope were similar for all models.

Case study 2

The detailed results are shown in Appendix D. In short, among 16,280 patients from 14 countries, 2,745 (16.9%) died due to any CVD related conditions. Using bootstrap validation, we found that the c-statistic ranged from 0.65 to 0.70 with good reproducibility, and that models with more predictors discriminated better. However, results of IECV indicate that the inclusion of additional predictors increased the heterogeneity in discrimination performance. Results of both bootstrap validation and IECV also indicate that inclusion of non-linear terms and/or interaction effects) did not improve discrimination performance. In calibration performance, the effect of complex modelling strategies was small in both summary estimates of O:E ratio and calibration slope and their generalizability.

4. Discussion

We illustrated how evidence synthesis methods can be used to evaluate the need of complex strategies for developing generalizable clinical prediction models in large clustered datasets. To this end, we applied IECV and quantified the model's average performance as well as its variability between clusters. In contrast to traditional internal validation methods, a major advantage of using IECV in large clustered data is that the external validity of prediction models can be assessed on multiple occasions, thereby allowing researchers to explore the generalizability of different modelling strategies directly during the development process.

In the case study 1, we found that adopting complex modelling strategies did not much improve the external validity of developed prediction models for HF. In particular, prediction models that were based on four commonly available variables yielded a c-statistic of 0.79, which is comparable to existing models for HF using even more than 10 predictors including laboratory tests [10, 11]. Although the inclusion of additional predictors marginally improved the discriminative performance, it also slightly increased the between-practice heterogeneity. When investigating model calibration, we found that all prediction modelling strategies yielded adequate calibration performance on average. However, because of between-practice heterogeneity, local revisions were often deemed necessary. In the case study 2, we also found that complex modelling did not meaningfully improve the generalizability of the prediction models, although the inclusion of additional predictors moderately improved their discrimination performance.

As we found in the case study 1, the incremental value of candidate predictors is often small in prediction model studies for the incidence of CVD [26, 27]. For instance, systematic reviews have demonstrated a lack of incremental value for cholesterol level [27], BMI [27], and even biomarkers (e.g., triglycerides, C-reactive protein) for predicting CVD [26]. For this reason, it may sometimes be more advantageous to consider the inclusion of non-linear effects or interaction terms, rather than adding more predictors. This strategy is common in machine learning, where methods no longer assume additive linear effects and adopt penalization to avoid overfitting. We mimicked the use of flexible modelling strategies by including non-linear effects and non-linear interaction terms. However, this strategy also failed to improve model discrimination. Similar findings also have been reported in prediction model studies for the prognosis of patients with CVD [28, 29]. For instance, a recent study adopting advanced machine learning algorithms failed to outperform traditional prediction models for readmissions in patients with HF, and yielded c-statistics around 0.60 [28]. In another study, discrimination performance to predict all-cause mortality in patients with coronary artery disease marginally increased from 0.793 (Cox regression model with 27 predictors) to 0.797 (random survival forests with 98 predictors) and to 0.801 (elastic net Cox regression model with 586 predictors) [29]. More generally, there is limited evidence that machine learning models can outperform simple prediction models involving additive linear terms, especially when predictions are only based on structured epidemiological data [30].

The following limitations need to be considered. In the first case study, the substantial presence of missing data was an important concern. Although we focused on the inclusion of variables with relatively few missing values, some were missing for more than 70% of participants. Multiple imputation is generally recommended to obtain reliable standard errors of the performance measures but only single imputation was pursued due to limited hardware processing capacity. There is still limited guidance on how to implement multiple imputation when developing and validating a prediction model in large clustered datasets. Key issues that remain unclear are (i) how to combine multiple imputation with sampling procedures (e.g., IECV) [31, 32], (ii) the order of pooling estimates (across imputations or across clusters first) [33]. Another limitation was that we were not able to include non-linear and interaction terms in the imputation model due to non-convergence issues. Therefore, continuous variables were imputed as a linear term and no interaction term was included in imputation models. This strategy may have favored simpler modelling strategies in IECV. For this reason, we implemented those modelling strategies in the case study 2 where the presence of missing data was much less a concern. And we found similar findings to those in the case study 1.

Second, eligible individuals in both case studies were enrolled more than ten years ago. It is possible that population characteristics have substantially changed over time, and that complex associations (e.g., non-linear predictor effects or interaction terms) have become more common.

Third, we focused on regression-based methods and did not evaluate other flexible modelling strategies such as neural networks or random forests. It is possible that these strategies could yield more promising results, especially if (interaction between) predictor effects cannot adequately be described using the regression-based methods considered here.

5. Conclusion

We recommend the use of IECV in large clustered datasets to assess the generalizability of prediction models during their development, and to identify whether complex modelling strategies may offer any advantages. In contrast to traditional internal validation methods, IECV allows to evaluate model performance in non-random hold-out samples with individuals from different settings or populations. In our case studies, we found that accurate prediction does not necessarily require complex modelling strategies, and that the need for local updating may be inevitable regardless of how much data are at hand during the model's development.

CRediT authorship contribution statement

Toshihiko Takada: Conceptualization, Formal analysis, Writing – original draft, Writing – review & editing. **Steven Nijman:** Formal analysis, Writing – original draft, Writing – review & editing. **Spiros Denaxas:** Data curation, Project administration, Resources, Writing – review & editing. **Kym I.E. Snell:** Methodology, Writing – review & editing. **Alicia Uijl:** Data curation, Writing – review & editing. **Tri-Long Nguyen:** Methodology, Writing – review & editing. **Folkert W. Asselbergs:** Project administration, Resources, Writing – review & editing. **Thomas P.A. Debray:** Conceptualization, Formal analysis, Supervision, Writing – original draft, Writing – review & editing.

Data sharing

Due to privacy laws and the data user agreement between the University College London and Clinical Practice Research Datalink, authors are not authorised to share individual patient data from these electronic health records. Requests to access data provided by Clinical Practice Research Datalink (CPRD) should be sent to the Independent Scientific Advisory Committee (ISAC) (https://www.cprd. com/ISAC/). The CALIBER portal (https://www. caliberresearch.org/portal/) does offer open sharing of phenotypic and analytic algorithms for use by other researchers. Aggregate data (e.g., model formulas, performance estimates) are available on request.

Funding

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking BigData@Heart grant no. 116074. TT was supported by the Uehara Memorial Foundation between August 2018 and July 2019. SN is supported by a Public-Private Study grant of the Dutch Heart foundation (2018B006). KIES is supported by a National Institute for Health Research School for Primary Care Research (NIHR SPCR) Launching Fellowship. FWA is supported by UCL Hospitals NIHR Biomedical Research Centre. TPAD is supported by the Netherlands Organisation for Health Research and Development (91617050), and by the European Union’s Horizon 2020 research and innovation programme under ReCoDID grant agreement no. 825746.

The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care.

Acknowledgements

This study is based in part on data from the Clinical Practice Research Datalink obtained under licence from the UK Medicines and Healthcare products Regulatory Agency. The data is provided by patients and collected by the NHS as part of their care and support. The interpretation and conclusions contained in this study are those of the author/s alone. Copyright © (2020), re-used with the permission of The Health & Social Care Information Centre. All rights reserved.

References

[1] Reilly BM, Evans AT. Translating clinical research into clinical practice: impact of using prediction rules to make decisions. Ann Intern Med 2006;144:201-9.

[2] Debray TP, Moons KG, Ahmed I, Koffijberg H, Riley RD. A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. Stat Med 2013;32:3158-80.

[3] Debray TP, Riley RD, Rovers MM, Reitsma JB, Moons KG, Cochrane IPDM-aMg. Individual participant data (IPD) meta-analyses of diagnostic and prognostic modeling studies: guidance on their use. PLoS Med 2015;12:e1001886.

[4] Debray TPA, de Jong VMT, Moons KGM, Riley RD. Evidence synthesis in prognosis research. Diagn Progn Res 2019;3:13.

[5] Riley RD, Ensor J, Snell KI, Debray TP, Altman DG, Moons KG, et al. External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: opportunities and challenges. BMJ 2016;353:i3140.

[6] Ahmed I, Debray TP, Moons KG, Riley RD. Developing and validating risk prediction models in an individual participant data meta-analysis. BMC Med Res Methodol 2014;14:3.

[7] Debray TP, Damen JA, Riley RD, Snell K, Reitsma JB, Hooft L, et al. A framework for meta-analysis of prediction model studies with binary and time-to-event outcomes. Stat Methods Med Res 2019;28:2768-86.

[8] Snell KI, Hua H, Debray TP, Ensor J, Look MP, Moons KG, et al. Multivariate meta-analysis of individual participant data helped externally validate the performance and implementation of a prediction model. J Clin Epidemiol 2016;69:40-50.

[9] Damen JA, Hooft L, Schuit E, Debray TP, Collins GS, Tzoulaki I, et al. Prediction models for cardiovascular disease risk in the general population: systematic review. BMJ 2016;353:i2416.

[10] Agarwal SK, Chambless LE, Ballantyne CM, Astor B, Bertoni AG, Chang PP, et al. Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study. Circ Heart Fail 2012;5:422-9.

[11] Hippisley-Cox J, Coupland C. Development and validation of risk prediction equations to estimate future risk of heart failure in patients with diabetes: a prospective cohort study. BMJ Open 2015;5:e008503.

[12] Smith JG, Newton-Cheh C, Almgren P, Struck J, Morgenthaler NG, Bergmann A, et al. Assessment of conventional cardiovascular risk factors and multiple biomarkers for the prediction of incident heart failure and atrial fibrillation. J Am Coll Cardiol 2010;56:1712-9.

[13] Denaxas S, Gonzalez-Izquierdo A, Direk K, Fitzpatrick NK, Fatemifar G, Banerjee A, et al. UK phenomics platform for developing and validating electronic health record phenotypes: CALIBER. J Am Med Inform Assoc 2019;26:1545-59.

[14] Denaxas SC, George J, Herrett E, Shah AD, Kalra D, Hingorani AD, et al. Data resource profile: cardiovascular disease research using linked bespoke studies and electronic health records (CALIBER). Int J Epidemiol 2012;41:1625-38.

[15] Uijl A, Koudstaal S, Direk K, Denaxas S, Groenwold RHH, Banerjee A, et al. Risk factors for incident heart failure in age- and sex-specific strata: a population-based cohort using linked electronic health records. Eur J Heart Fail 2019;21:1197-206.

[16] Yang H, Negishi K, Otahal P, Marwick TH. Clinical prediction of incident heart failure risk: a systematic review and meta-analysis. Open Heart 2015;2:e000222.

[17] The English Indices of Deprivation 2019. https://www.gov.uk/government/statistics/english-indices-of-deprivation-2019. Date accessed: October 19, 2020.

[18] CALIBER. https://www.ucl.ac.uk/health-informatics/caliber. Date accessed: October 19, 2020.

[19] Resche-Rigon M, White IR. Multiple imputation by chained equations for systematically and sporadically missing multilevel data. Stat Methods Med Res 2018;27:1634-49.

[20] Simon N, Friedman J, Hastie T, Tibshirani R. Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent. J Stat Softw 2011;39:1-13.

[21] Debray TP, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KG. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol 2015;68:279-89.

[22] Prognosis research in health care : concepts, methods, and impact. 1st edition. ed. NewYork, NY: Oxford University Press; 2019.

[23] Debray TP, Damen JA, Snell KI, Ensor J, Hooft L, Reitsma JB, et al. A guide to systematic review and meta-analysis of prediction model performance. BMJ 2017;356:i6460.

[24] Snell KI, Ensor J, Debray TP, Moons KG, Riley RD. Meta-analysis of prediction model performance across multiple studies: Which scale helps ensure between-study normality for the C-statistic and calibration measures? Stat Methods Med Res 2018;27:3505-22.

[25] Sandercock PA, Niewada M, Czlonkowska A, International Stroke Trial Collaborative G. The International Stroke Trial database. Trials 2011;12:101.

[26] Ioannidis JP, Tzoulaki I. Minimal and null predictive effects for the most popular blood biomarkers of cardiovascular disease. Circ Res 2012;110:658-62.

[27] van Bussel EF, Hoevenaar-Blom MP, Poortvliet RKE, Gussekloo J, van Dalen JW, van Gool WA, et al. Predictive value of traditional risk factors for cardiovascular disease in older people: A systematic review. Prev Med 2020;132:105986.

[28] Frizzell JD, Liang L, Schulte PJ, Yancy CW, Heidenreich PA, Hernandez AF, et al. Prediction of 30-Day All-Cause Readmissions in Patients Hospitalized for Heart Failure: Comparison of Machine Learning and Other Statistical Approaches. JAMA Cardiol 2017;2:204-9.

[29] Steele AJ, Denaxas SC, Shah AD, Hemingway H, Luscombe NM. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. PLoS One 2018;13:e0202344.

[30] Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J Clin Epidemiol 2019;110:12-22.

[31] Wood AM, Royston P, White IR. The estimation and use of predictions for the assessment of model performance using large samples with multiply imputed data. Biom J 2015;57:614-32.

[32] Mertens BJA, Banzato E, de Wreede LC. Construction and assessment of prediction rules for binary outcome in the presence of missing predictor data using multiple imputation

and cross-validation: Methodological approach and data-based evaluation. Biom J 2020;62:724-41.

[33] Burgess S, White IR, Resche-Rigon M, Wood AM. Combining multiple imputation and meta-analysis with individual participant data. Stat Med 2013;32:4499-514.