a model not a prophet

operationalising patient-level prediction
using observational data networks

ross d. williams

A Model, Not a Prophet: Operationalising patient-level prediction using observational data networks

Een model, geen profeet: patient-level prediction operationaliseren met gebruik maken van observationeeldata netwerken

Ross D. Williams

ISBN: 978-94-6361-896-0

Cover design: Eve Nightingale
Layout and printing: Optima Grafische Communicatie, Rotterdam | www.ogc.nl

A Model, Not a Prophet: Operationalising patient-level prediction using observational data networks

Een model, geen profeet: patient-level prediction operationaliseren met gebruik maken van observationeeldata netwerken

Proefschrift

ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rotterdam
op gezag van de
rector magnificus

Prof.dr. A.L. Bredenoord

en volgens besluit van het College voor Promoties.
De openbare verdediging zal plaatsvinden op

Dinsdag 26 September 2023 om 13:00 uur

door

Ross David Williams
geboren te Londen, Verenigd Koninkrijk.

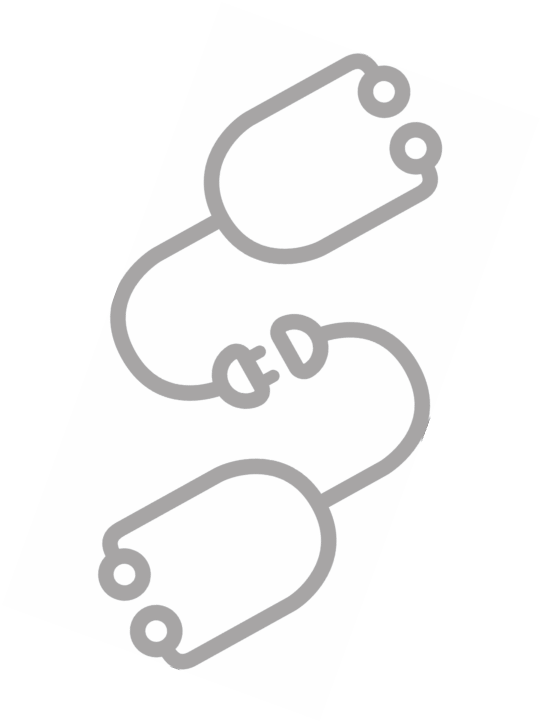
Erasmus University Rotterdam

Promotiecommissie

Promotoren	Prof.dr.ir. P.R. Rijnbeek
	Prof.dr. E.W. Steyerberg
Overige leden	Prof.dr. P.J.E Bindels
	Prof.dr. H.F. Lingsma
	Prof.dr. S le Cessie
Copromotoren	Dr.ir. D. van Klaveren
	Dr. J.M. Reps

For Powell, for inspiring a lifetime of investigation

TABLE OF CONTENTS



General Introduction

Ross D. Williams

Predicting the future is hard, but the desire to do so has been prevalent throughout human history. Whether this is the Celtic druids using bird flight paths to make predictions about the weather, or the Pythia prophesising at Delphi, humans have been interested in knowing what is going to happen in the future. The same is true today, although the methods of predicting are arguably more sophisticated. The central idea remains the same: If I can know what is going to happen using the information I have now, I can intervene now to change the future. Considering the current "Big Data" era, prediction is everywhere, from the next word suggested when you type an email, to whether you get a mortgage or what you will buy next. Healthcare is also increasingly opening to the possibilities that prediction modelling will bring, specifically predictions for individual patients and the prospects this brings to personalise their care. Predictive analytics have the potential to revolutionise healthcare by increasing the level of personalisation across the spectrum of diseases and treatments in a way previously impossible. To do this we need to have high performing, trustworthy patient-level prediction (PLP) models. Such models could inform healthcare professionals of who is at high risk and this can be used to target these patients for closer monitoring, prescribing different medications or a myriad of other options for treatment personalisation. The models themselves need to be clearly and effectively communicated; it does not matter if you have a perfect model, if no one can find it or use it then it may as well not exist. Currently this communication and model dissemination is lacking, either the model is not provided, or the results are ambiguously reported. The ideal situation is an open reporting system that empowers a community to find, share and contribute to the evidence process.

Within healthcare, knowing a patient's risk, or knowing about some future event can aid in the development of a patient's treatment pathway. Let us consider an example. The American Diabetes Association (ADA) guidelines have been tending towards a more personalised treatment pathway over the past 5 years. The 2017 (1) and 2018 (2) ADA guidelines included recommendations that at initial diabetes diagnosis, the patient should receive a treatment with metformin and lifestyle interventions. If the patient's HbA1C level remains above 9% after three months then a further drug should be added and this should be assessed based upon the risk profiles of the drugs which best suit the patient. In the 2019 guidelines (3) the advice was updated to stratify patients based upon established heart failure (HF), atherosclerotic cardiovascular disease and chronic kidney disease. From this a clear trend is emerging in the desire to stratify and personalise treatments. If a patient is at risk of HF, then a diabetes treatment with diuretic effects (e.g. sodium-glucose cotransporter 2 inhibitors) is known to be beneficial, but if a patient were to take a medication (e.g. thiazolidinediones) which has water retention as a known side effect this could be detrimental. It would be beneficial for the patient population here to stratify treatment by risk. Then the question becomes what are different risks for the patient, what is most concerning, and where can the most benefit be gained. It should be noted that the ADA guidelines do not in fact provide any recommendation for how this risk should be assessed, just that it should be. By creating risk prediction models for this problem it is clear that these can help in decision making to impact patient care.

When treating a patient according to the aforementioned guidelines, we can expect to encounter the following situation. A patient initiating a pharmaceutical intervention for diabetes additional to metformin, has multiple options. How to choose between these options is a complex question. When population-level effects for the different medications are similar, the selection is traditionally decided by physician experience or non-patent centric factors such as cost and availability. However, each drug and each patient has a unique risk profile and as such on an individual level the reaction of a patient and the effectiveness of the treatment can vary wildly. This is where a PLP model can have an impact. If we can *predict* the risk of an outcome, then we can match that risk to beneficial effects of a drug (or avoid drugs with a known compounding negative effect). If a patient's risk of heart failure is known, then their treatment can be altered accordingly. Without a PLP model only averages for the population can be given, but how relevant would this be to the individual patient? This approach suggests that a patient aged 65 with no comorbidities has the same risk of heart failure as a patient aged 90 with fluid retention and high blood pressure. Clearly this is incorrect but we have limited information to suggest what the difference in outcome risk would be between these patients, other than an expectation that the risk is different. Another approach is for the physician to draw on their experience and estimate what the risk is, but this is unreproducible, unverifiable, and susceptible to bias. The number of patients a doctor will treat in their career lies also in the thousands, but there are millions of patients in EHRs now. Failing to utilize this vast collection of information would be a missed opportunity to attempt to improve patient care.

Part I of this thesis concerns itself with multiple aspects of PLP modelling, including technical challenges in the development and validation of PLP models, and how best to disseminate the results of studies in a manner which facilitates their access, understanding, and validation. Part II will cover the development and validation of PLP models for specific clinical applications, according to the best practices described in part I of this thesis and literature.

PART I: PATIENT-LEVEL PREDICTION MODELLING

What is PLP modelling, and why is it interesting? Each time a decision is made in healthcare, it is a question of assessing the risks of treatment side effects, the severity of these side effects, the beneficial effects of treatment and how these combine in the patient pathway. Should I give a patient an intervention I know has side effects? Who should be called in for monitoring and potentially invasive screening? Who will benefit the most from a treatment in a resource-restricted environment? All of these involve some form of risk assessment for the patient. The underlying idea is to assess the risks, the costs of the potential side effects and the expected benefit of the treatment. With the generation of massive amounts of patient data, there is an opportunity to do personalised risk assessment in a consistent, reproducible and open-science manner.

This idea of personalisation is key to the future of healthcare. As more and more treatments become available, it is likely many of these will be similarly beneficial valuable when considered at population level, but can have different impacts for different patients. As such there is a huge potential gain from identifying who will respond best to what medication, and risk plays an important role in that.

The potential gains that could be made in personalisation of healthcare through the use of PLP modelling has lead to a dramatic increase in the number of models developed each year. However, there is a wide diversity in the approaches taken for building prediction models and this has led to models of varying quality being produced. Fortunately, there do exist best practices that if followed can produce models of clinical impact. Building a prediction model consists of multiple steps and can be best performed with multiple interdisciplinary stakeholders. Reps et al. produced a framework for developing optimal prediction models in a transparent process producing results according to open science principles(4-6). This framework is made possible through the standardisation of health data to the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) maintained by the Observational Health Data Sciences and Informatics (OHDSI) community (see Figure 1)(7).

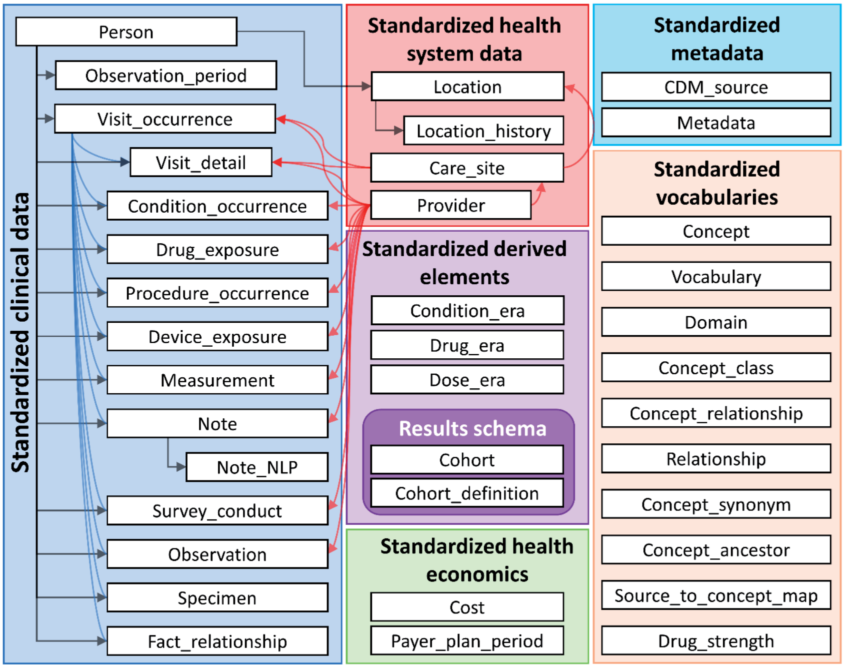


Figure 1 The structure of the OMOP-CDM

The OMOP-CDM aims to improve both the syntactic and semantic interoperability of the health data. Standardising the clinical data to a common format as shown in Figure 1 (blue box), enables the use of standardised analysis pipelines such as the PLP framework. The use of standardised vocabularies (orange box) improves the semantic interoperability, i.e., it facilitates the identification of clinical concepts using a common terminology. More information about the OMOP-CDM can be found in The Book of OHDSI (https://book.ohdsi.org).

As shown in Figure 2, a prediction problem can be defined by the target and outcome cohorts (target cohort being patients we want to make a prediction for, outcome being what we want to predict) and the time at risk. These three elements, along with the look back period, give the minimum requirement for correctly specifying a prediction problem. The index date of the target cohort should by defined as the moment the decision is made in clinical practice. For example, a model that is intended to inform the decision making of which secondary medication to choose for a diabetes patient, should have as index date the date when a secondary medication is started and not the date of diagnosis or initiation of metformin treatment. Similar care should be taken when choosing outcomes and time at risk. The time at risk should use a sensible period of time during which it is expected there is some relationship between the target and the outcome. For example, a time at risk of 30 days for an oncological outcome is unlikely to be sensible, whereas 1-5 years could be informative.

The method described above can be abstracted into a prediction framework. Changing the target and outcome cohorts and the time at risk would allow for the same mechanism to be used for different prediction problems. The use of such a framework removes several key barriers that have prevented predictive analytics from impacting healthcare. For example, the availability of the developed models, the cohorts and model settings required to implement the models, and clear instructions on best practices for validation. All of these are included

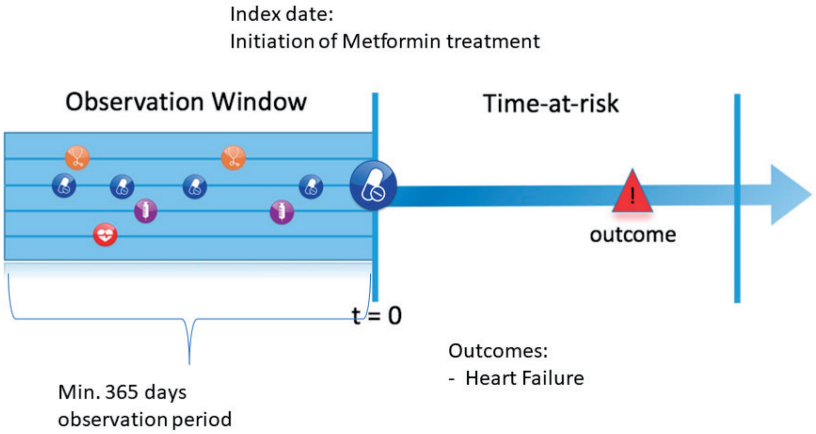


Figure 2 Detailing the prediction specification for heart failure occurring within 365 days of initialising metformin as an anti-diabetic treatment

as standard in software packages created in the research pipeline of Reps et al. There remain however multiple challenges to the implementation of prediction models in clinical practice. These challenges include a lack of trust in the models, a lack of understanding of the models and modelling process, difficulties in implementing models (often due to large numbers of covariates) and issues with the reporting of models, which often do not follow open science principles. Chapter 2 will address the issue of complex models and will demonstrate a solution. Further issues of trust will be addressed in chapter 4 by considering how performance should be interpreted in terms of the model complexity (how much better is a more complex model) and database adjusted performance expectation (a lower performance does not always mean a model is bad, it could be the database is harder to predict in). Chapter 5 will address the issues of openness, reporting and model availability.

Internal Validation

Demonstrating good performance is often a difficult thing to do. We could consider a model to be performing "well" if a patient's treatment pathway is improved because of a change in treatment decision based on the model. Clearly this is difficult to demonstrate, for a multitude of reasons including the lack of a counter-factual (i.e. we cannot know what would have happened in the situation where the model was not used). Further, we need some way of deciding if the model works or not before it can be used in clinical practice.

This leads us to the question of what a good model is and how can this be demonstrated? When considering what is important for a prediction, one of the factors is "if I make a prediction does it come true?" This can be reframed in the medical context as "of all the patients who are treated, some will go on to experience the outcome. Can they correctly be distinguished from the others who will not experience it?" This brings us to the concept of *discrimination.* There are many metrics measuring discrimination, for example, sensitivity: *of patients who will get the outcome, what proportion do I correctly predict as getting the outcome?* and specificity: *of patients who will not get the outcome, what proportion do I correctly predict not to?* These metrics are derived from a confusion matrix (Table 1). Fundamentally there are true positives and true negatives (correct predictions) and false positives and false negatives (incorrect predictions). Importantly, in the case of a regression model, a probability threshold must be set to determine what counts as a predicted positive or negative, i.e., everyone above the threshold will be predicted to get the disease and everyone below to not get it. Picking this threshold is going to be very specific to the problem and the specific implementation. Different problems give different costs to a false positive and a false negative. Consider the case of a highly infectious disease. If persons

Table 1 Definition of a confusion matrix showing what are true/false positives and negatives

	Ground Truth Positive	Ground Truth Negative
Predicted Positive	True Positive	False Positive
Predicted Negative	False Negative	True Negative

receive a false positive test then the cost is that they will be worried for their health and will likely have to isolate for a period of time (or indeed take a new test, which in this case could be a nasal swab with mild discomfort). However, a false negative would mean they were then going to continue on as normal and likely spread the disease further to a group of people. In this situation the false negative has a higher associated cost than a false positive. However, in another situation the false positive could lead to further testing which might not be as benign as a nasal swab. It could lead to a brain biopsy for example, which is a much more involved surgical process and as such the cost here of the false positive increases. Given the nuances here it becomes useful to develop a metric that is threshold independent. This is where the concept of the receiver operator characteristic (ROC) curve and the area under the ROC curve (AUC) comes in.

An example of an ROC curve is presented in Error! Reference source not found.. This curve makes it possible to view the trade-off between sensitivity and specificity at the various possible thresholds. If the AUC is calculated then this provides a more general view of the performance of a model. AUC ranges (generally) between 0.5 and 1, where 1 indicates perfect discrimination and 0.5 represents random chance. A useful way of thinking about this is to take the example of two randomly picked patients. One of these patients will go on to experience the outcome of interest and the other will not. A model with an AUC of 1 will always give the patient who experiences the outcome a higher chance than the patient who does not. A model that gives an

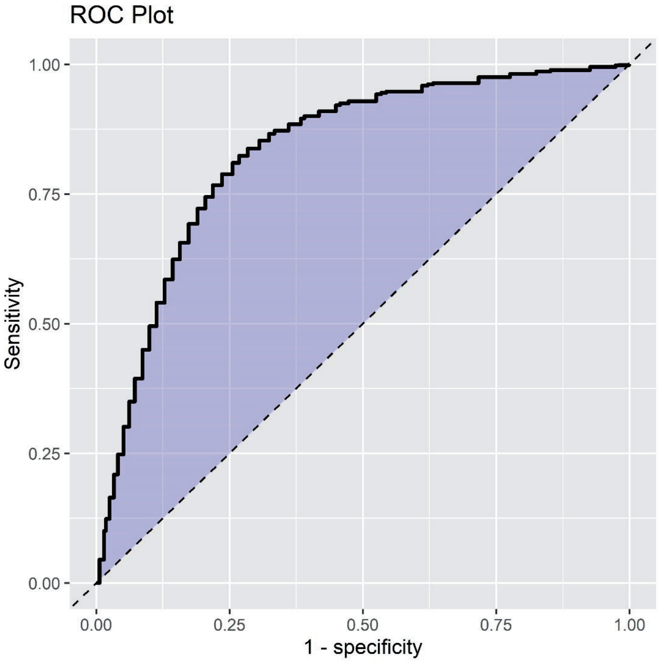


Figure 3 Example of an ROC plot

AUC of 0.5 will then give this order half the time (equivalent to a coin toss). It is conceivable to have an AUC of less than 0.5, but then we could simply invert the suggestion to achieve an AUC of greater than 0.5 and therefore the AUC is generally considered to range between 0.5 and 1.

Another important way of considering performance is to determine whether the absolute risk assigned by the model is correct. If a patient receives a 25% chance of developing the outcome, considering this patient split into 4 parallel universes this risk would be "correct" if they were to experience the outcome in 1 of these universes. Unfortunately, we do not have access to parallel universes, but we could consider 4 patients, each of whom receive a predicted risk of 25%, if one of them goes on to experience the outcome then we would consider this risk to be "correct". This is known as calibration(8, 9). If a model is well-calibrated, then the observed fraction of patients who experience the outcome is equal to the fraction predicted to experience it. There are various methods of exploring this which give insight into various calibration properties. Van Calster et al. excellently describe several of these metrics(10), such as calibration in the large which compares the mean predicted risk to the outcome occurrence. However these metrics often lack a nuance required for calibration. Often the calibration of a model can vary across the predicted risks. For example, the model could be well calibrated for a predicted risk of between 0-10% but poorly calibrated for a predicted risk of 40-50%. Static metrics that assess the calibration will miss this nuance. A common method of calibration assessment is to look at calibration curves(11).

An example of these is given in Figure 4 Here the predicted risk is plotted against the observed outcomes by fitting a LOESS model. The advantage of this over other calibration metrics is to show the calibration performance over the range of predicted probabilities and then we can see that for different risk ranges (e.g. low, medium and high) whether the model

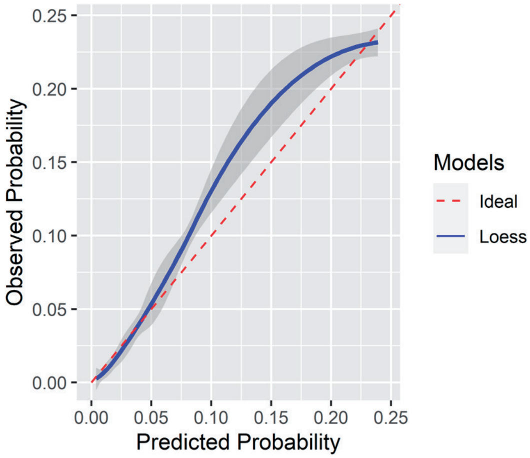


Figure 4 A smooth Calibration plot also including the absolute distribution of outcomes. Here Loess is the smoothed function to show the calibration of the model, the ideal line shows perfect calibration

is well or poorly calibrated. There are lots of ways of displaying calibration plots, smoothed as seen in the figure here (the authors personal favourite), but the use of splines and even linear fit models are acceptable. The smooth plot trades off some simplicity with more flexibility to show different miscalibration in different areas.

Internal validation is essential for model development, however it is not enough to satisfy requirements for the deployment of models in clinical practice. Internal validation provides evidence of how one model performs in one setting in one database. In order to generalise this to use the model in wider care more extensive *external* validation must be performed.

External Validation

Increasingly, external validation is being seen as essential to the understanding, trust in, and implementation of PLP models(12-14). Primarily, external validation is used to assess how well a model developed in one setting transports to another. An example could be testing a model developed in a claims database in another claims database, or it could be examining whether a model developed in a Dutch general practice (GP) database transports and performs well in a UK GP database. External validation can also mean applying the model in a similar setting; this can include slight differences in the definitions of the target and outcome cohorts. External validation has traditionally been a time-consuming and complicated process(15). A major reason has historically been the lack of syntactic and semantic database interoperability between databases(16). This means that if researchers want to validate a model in new data, they need to create new target and outcome cohorts, understand how to map covariates and understand any differences in the representation of patients between the original and new coding system. With the introduction of a common data model, many of the technical barriers have been removed. There remain however several questions surrounding model performance evaluation in this context:

Firstly, how should external validation be performed and what are the best practices for this? Given that there are multiple differing types of external validation, each assessing a different model characteristic it is essential to define what the validation is aiming to do(17). The simplest, and most common, form of external validation is to directly assess the model "generalisability". This means that a researcher applies the model on new data where the population matches to the original setting, e.g. , the target and outcome cohort definitions, and the time at risk are identical. Another method can be to examine the "transportability" of the model. This involves applying the model on new data with a similar population. Chapter 3 of this thesis will address this issue, and methods of assessing external validation in a timely fashion in a federated network of databases. This chapter will discuss a framework for performing external validation and how this process can be done systematically and efficiently within the OHDSI ecosystem.

Secondly, the technical challenge of externally validating a model, is only one facet of the evidence generating process. When externally validating a model, performance must be given a context. Context means that the performance can be measured against a reference. Training a model locally in a database, sets this expectation by providing a reference or benchmark. If a performance drop is found when *externally validating* a model compared to the result of the *internal validation*, then this could be because the model was tuned too much to the training data to properly transport to unseen data, i.e. the model was overfit or it needs recalibration. However, it could also be that the performance achieved is similar to the performance of a model that is trained on that same database. In other words, the model performs as well as possible in the context of the available data in that database. We need a model development approach that provides this context. Furthermore, simpler models are preferred as they are more easily clinically implemented and as such understanding the performance gain compared to the baseline of using only age and sex is valuable to contextualize the performance of the more complex model. Training a baseline model (most commonly age and sex) is essential to understanding what the performance gains are in relation to the increasing model complexity and difficulty of clinical implementation. A model with a minor performance increase but hundreds of covariates is probably not the best candidate to be used in clinical practice.

Thirdly, given the interoperability of these databases, there is an opportunity to use them not only to perform external validation, but to improve the internal performance of the model in the development phase. Current best practice is to develop a model in a single database and then perform validations in other databases. However, it is also possible to use data from multiple databases at the same time to develop a single model. This could result in a better performing model and improved generalisability and transportability. Pooling multiple databases and then building prediction models, however, is often not feasible due to strict governance rules and patient privacy(18, 19). However, it is possible to use a privacy-by-design approach by building an ensemble classifier using a federated network of databases, i.e., the data stays local within their safe haven, we bring the tools to the data, and only share the results. Ensemble learning is the process of producing multiple models, potentially pruning the set of models and then combining the remaining models. Often the ensemble increases model performance and stability compared to any single classifier. Ensembles either combine homogeneous models (same learning algorithm) or heterogeneous models (different learning algorithms). Homogeneous ensembles use the same learning algorithm but modify the perspective by using different training data (e.g., different instances, different features or by adding noise), different metrics or using different model settings (e.g., hyper-parameter values). Heterogeneous models take a different perspective as each learning algorithm makes different assumptions about the data. This will be explored in Chapter 1.

Finally, prediction models containing many predictors could have a higher performance than those with only a small subset of predictors, but there are clear issues with implementing such a model with many predictors in clinical practice. As a rough estimate, models with more than 25 predictors start to become challenging when considering the time-restricted environments within which the models need to be applied. If the desire is for the model to have immediate clinical utility and impact, it is necessary to find some mechanism to develop parsimonious models, or to develop a process of parsimonisation of more complex models. It is possible to develop a tool that directly integrates into an EHR system. However, this is a complicated process and would likely involve multiple rounds of review and software development, slowing down the speed at which the model can find itself in the clinic and increasing the cost. In a time-sensitive situation, for example at the start of a pandemic, a rapid response is likely to have the largest impact. This raises another question which is, for a novel disease or for a disease which is rare, what is the best way to develop a model? There has been a large body of research which attempts to use complex modelling and multiple steps to develop a model in a data-restricted situation. However, why should we necessarily start from nothing? Given insufficient data to effectively train and validate a model, a smarter solution needs to be found. This is where proxy learning can help. By doing an initial training step in a less restricted environment, for example using a disease that has a large cohort of patients, initial training can be performed and then this can be *externally validated* in the true target cohort to assess if performance transfers. Then the model can be updated or recalibrated as needed. In doing this the data used from the disease of interest can be kept for performance evaluation which then gives more weight to the evidence that is generated in this process. This process will be explored in detail in Chapter 2.

Dissemination

There also remains an unsolved problem of dissemination of results in a manner that is accessible, understandable and available to multiple stakeholders of varying degrees of expertise. There have been several excellent progressions in the reporting of models, particularly the introduction of the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) statement and accompanying checklist(20). This has lead to increased clarity in reporting and wider sharing of essential information needed to analyse the models developed. The TRIPOD statement is an attempt to change *what* is reported and now *how* it is reported. How a model is reported is usually through the publication of an article which contains (hopefully) the essential model information. This is static and does not allow for easy model exploration, use and updating. It also does not provide any flexibility on the part of the reader to explore the model and its performance. If we turn our focus to models developed using the standardised pipeline provided by the PLP package and the OMOP CDM, the standardisation means that the modelling process is consistent at every step, including the production of a standard results object (this means that it does not matter what specific algorithm a researcher uses, or their problem setting, the results object will look identical e.g. the performance scores will be located in identical places). This standardisation gives the opportunity to create an application that will allow exploration of these results across the data network.

An excellent resource for the assessment of prediction models is the Tufts PACE clinical prediction model registry(21). This is an intuitive and easy to access collection of clinical prediction models, in this thesis referred to as patient-level prediction models. It provides the model and often information on the performance and validations of the models as well as published articles about the models. The registry itself includes models from diverse sources and as such there is no underlying standardisation. This lack of standardisation means the registry has some limitations, including displaying models in a common format, facilitating the external validation of these models, and the static display of evaluation statistics limiting the flexibility for the user.

A centralised, interactive PLP model library would provide researchers, regulators and clinicians with a user-friendly experience to discover and evaluate prediction models. By providing this, along with the ability to download models (and the relevant cohorts and parameters needed to externally validate a model), upload validations and access relevant articles, many barriers to the adoption of PLP models in practice can be removed. By providing the results in this dynamic format, stakeholders will not be constrained to the traditional static methods of model review, e.g. assessing the limited results provided in a journal article. They will have direct access to the aggregated results and graphics and as such have more flexibility to explore what they find important or examine data that allows them to address or confirm misgivings that they have. By providing researchers with the ability to find these models and to download all the necessary components (target and outcome cohorts, full model) to run a validation, it is hoped that the rate at which models will be externally validated by independent researchers will increase. The current status of dissemination is discussed and an improvement on this is developed and demonstrated in chapter 5.

PART II: CLINICAL APPLICATIONS

The second part of this thesis details two prediction models. These models can be seen as intertwined with the research in part 1. They use some of the best practices developed but they also generated ideas for some of the research that would be conducted.

The first of these models concerns predicting short-term mortality after a total knee replacement (TKR)(22). TKR is a safe and cost-effective surgical procedure for treating severe knee osteoarthritis. Although complications following surgery are rare, prediction tools could help identify high-risk patients who could be targeted with preventative interventions. By creating a risk stratification tool to target this outcome, patients can be better informed about the risks and decide together with their clinician on whether to proceed with the surgery. Specifically a parsimonious model that could be easily distributed and quickly analysed could support clinicians in shared decision making and risk assessments when deciding on surgical interventions, as well as aiding in targeting preventative treatment. This study is discussed in detail in chapter 6.

The second concerns predicting adverse health outcomes amongst rheumatoid arthritis (RA) patients. Compared to the general population, patients with RA have an increased risk of treatment-related adverse events(23). Identification of RA patients at high risk of adverse health outcomes remains a major challenge (24, 25). Importantly, however, there are treatment options available to target these known comorbidities. Initiating methotrexate (MTX) monotherapy (with glucocorticoids) immediately post RA diagnosis (26, 27), is the most common treatment for RA globally. Using prediction models to evaluate patient-level risks in RA patients initiating first-line MTX monotherapy could allow clinicians to target those at high risk of adverse health outcomes for increased screening or monitoring throughout the course of treatment. Many prediction models have been developed for adverse health outcomes in RA patients, mostly focusing on the risk of either cardiovascular disease or serious infection (28-35). Importantly, none of these models have been subjected to extensive external validation, which is necessary to understand the performance of a prediction model (16). Further details of this process are discussed in chapter 7.

An additional clinical model was developed, although this chapter appears in part 1. At the onset of the COVID-19 pandemic, it was thought that prediction models could play a vital role in the risk assessment both at a patient and health authority/hospital administration level. One of the essential ideas behind the modelling in this early stage was to try to reduce the pressure on the health system. As such building a model from historical patient information and demographics (e.g. without requiring any new diagnostic tests) would be beneficial. This is because doing so allows for patients to be triaged off site, either by phone or videocall. This could potentially have a beneficial impact on the intensity of pressure on the healthcare system by reducing the patients seen at either primary or emergency care sites. To do this the COVID-19 Estimated Risk (COVER) scores were developed. These quantified a patient's risk of hospital admission with pneumonia (COVER-H), hospitalization with pneumonia requiring intensive services, or death (COVER-I), or fatality (COVER-F) in the 30-days following COVID-19 diagnosis using historical data. This modelling process used two interesting techniques to produce models rapidly and were easy to use. The first was the use of historical influenza data as a proxy for training the models to preserve the COVID data for testing the models, the second was the use of phenotype predictors in a process of parsimonisation to produce 9-predictor models. These could then easily be deployed. This work is detailed in chapter 2.

All 3 of these studies were produced in a research process called a study-a-thon. This is an intense method of performing epidemiological research where a team of researchers focus entirely on one project. This involves close cooperation with a large multidisciplinary team to symbiotically produce relevant questions, protocols, analysis plans, and to execute the research. By cooperating in this manner, the research phase of the projects can be performed within a week, compared to months that a traditional study takes. This allows for flexible and rapid production of evidence to guide treatment choices. The TKR and RA studies were performed in person and the COVID work was done remotely. Both mechanisms worked excellently, although it is certainly more enjoyable to do a study-a-thon in person in Barcelona than remotely during isolation.

Contributing to better healthcare outcomes is, was, and should always remain, the main goal of any research performed in the healthcare prediction modelling domain. It is possible to lose sight of this and to lose sight of the connection between the modelling process and the intricacies of machine learning. What the focus must always be on is that the research will at some point contribute to improving outcomes for real patients. These are people with families and lives and happiness, mostly they will not care if you use a deep learning model or a LASSO or bootstrapping or train test splits. They care about the quality and longevity of their lives. If strong evidence of performance of a model can be provided and it can demonstrably help them, then the patient will be satisfied.

On that note of providing evidence, we can look to the thoughts of Richard Feynman.

> *"If it [a theory] disagrees with experiment, it's wrong. In that simple statement is the key to science. It doesn't make any difference how beautiful your guess is, it doesn't make any difference how smart you are, who made the guess, or what [their] name is. If it disagrees with experiment, it's wrong. That's all there is to it."*
>
> — Prof. Richard P. Feynman, Lecture at Cornell University, 1964

Aims of the thesis:

1. To improve the development process of patient-level prediction models
2. To improve the validation of patient-level prediction models
3. To improve the dissemination of patient-level prediction models
4. To demonstrate how a best practice framework can be applied to specific clinical prediction problems.

BIBLIOGRAPHY

1. Marathe PH, Gao HX, Close KL. American Diabetes Association Standards of Medical Care in Diabetes 2017. J Diabetes. 2017;9(4):320-4.
2. Association American D. Updates to the Standards of Medical Care in Diabetes-2018. Diabetes Care. 2018;41(9):2045-7.
3. Care F. Standards of Medical Care in Diabetes 2019. Diabetes Care. 2019;42(Suppl 1):S124-S38.
4. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernandez-Bertolin S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed. 2021;211:106394.
5. Reps J, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. PatientLevelPrediction: Package for patient level prediction using data in the OMOP Common Data Model. 2018.
6. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
7. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.
8. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic t, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
9. Austin PC, Steyerberg EW. The Integrated Calibration Index (ICI) and related metrics for quantifying the calibration of logistic regression models. Statistics in medicine. 2019;38(21):4051-65.
10. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. J Clin Epidemiol. 2016.
11. Austin PC, Steyerberg EW. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Stat Med. 2014;33(3):517-35.
12. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022;29(5):983-9.
13. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7.
14. Steyerberg EW, Bleeker SE, Moll HA, Grobbee DE, Moons KG. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56(5):441-7.
15. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34.
16. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med. 2019;2:79.
17. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49-58.
18. Zerka F, Barakat S, Walsh S, Bogowicz M, Leijenaar RTH, Jochems A, et al. Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care. JCO Clin Cancer Inform. 2020;4:184-200.
19. Bogowicz M, Jochems A, Deist TM, Tanadini-Lang S, Huang SH, Chan B, et al. Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. Sci Rep. 2020;10(1):4542.

20. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.

21. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. Diagn Progn Res. 2017;1:20.

22. Williams RD, Reps JM, Group OEKA, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. Knee Surg Sports Traumatol Arthrosc. 2022;30(9):3068-75.

23. Yang C, Williams RD, Swerdel JN, Almeida JR, Brouwer ES, Burn E, et al. Development and external validation of prediction models for adverse health outcomes in rheumatoid arthritis: A multinational real-world cohort analysis. Semin Arthritis Rheum. 2022;56:152050.

24. Dougados M, Soubrier M, Antunez A, Balint P, Balsa A, Buch MH, et al. Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA). Ann Rheum Dis. 2014;73(1):62-8.

25. Turesson C. Comorbidity in rheumatoid arthritis. Swiss Med Wkly. 2016;146:w14290.

26. Singh JA, Saag KG, Bridges SL, Jr., Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol. 2016;68(1):1-26.

27. Smolen JS, Landewe RBM, Bijlsma JWJ, Burmester GR, Dougados M, Kerschbaumer A, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. Ann Rheum Dis. 2020.

28. Arts EE, Popa CD, Den Broeder AA, Donders R, Sandoo A, Toms T, et al. Prediction of cardiovascular risk in rheumatoid arthritis: performance of original and adapted SCORE algorithms. Ann Rheum Dis. 2016;75(4):674-80.

29. Crowson CS, Hoganson DD, Fitz-Gibbon PD, Matteson EL. Development and validation of a risk score for serious infection in patients with rheumatoid arthritis. Arthritis Rheum. 2012;64(9):2847-55.

30. Crowson CS, Rollefstad S, Kitas GD, van Riel PL, Gabriel SE, Semb AG, et al. Challenges of developing a cardiovascular risk calculator for patients with rheumatoid arthritis. PLoS One. 2017;12(3):e0174656.

31. Curtis JR, Xie F, Chen L, Muntner P, Grijalva CG, Spettell C, et al. Use of a disease risk score to compare serious infections associated with anti-tumor necrosis factor therapy among high- versus lower-risk rheumatoid arthritis patients. Arthritis Care Res (Hoboken). 2012;64(10):1480-9.

32. Curtis JR, Xie F, Crowson CS, Sasso EH, Hitraya E, Chin CL, et al. Derivation and internal validation of a multi-biomarker-based cardiovascular disease risk prediction score for rheumatoid arthritis patients. Arthritis Res Ther. 2020;22(1):282.

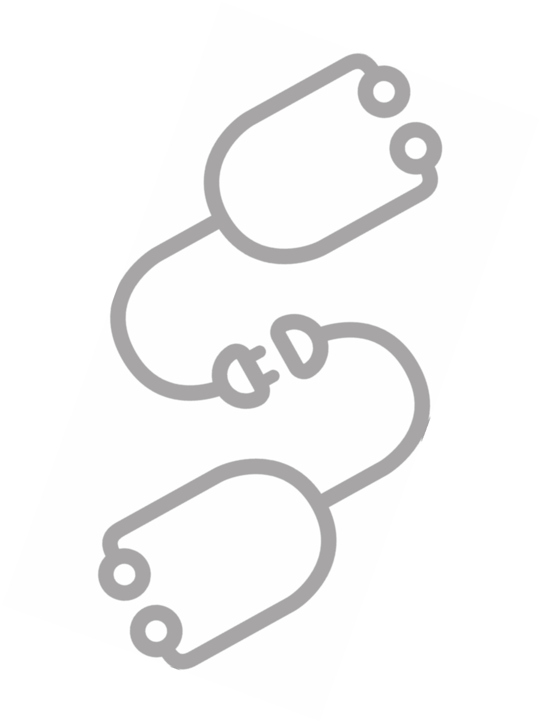
33. Solomon DH, Greenberg J, Curtis JR, Liu M, Farkouh ME, Tsao P, et al. Derivation and internal validation of an expanded cardiovascular risk prediction score for rheumatoid arthritis: a Consortium of Rheumatology Researchers of North America Registry Study. Arthritis Rheumatol. 2015;67(8):1995-2003.

34. Strangfeld A, Eveslage M, Schneider M, Bergerhausen HJ, Klopsch T, Zink A, et al. Treatment benefit or survival of the fittest: what drives the time-dependent decrease in serious infection rates under TNF inhibition and what does this imply for the individual patient? Ann Rheum Dis. 2011;70(11):1914-20.

35. Wang D, Yeo AL, Dendle C, Morton S, Morand E, Leech M. Severe infections remain common in a real-world rheumatoid arthritis cohort: A simple clinical model to predict infection risk. Eur J Rheumatol. 2020.

Part I

Methodological Research



Learning patient-level prediction models across multiple healthcare databases: Evaluation of ensembles for increasing model transportability

Jenna Marie Reps[1]*
Ross D. Williams[2]*
Martijn Schuemie[1]
Patrick Ryan[1]
Peter Rijnbeek[2]

[1]Janssen Research and Development, Raritan, NJ, USA;
[2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

* Co-first authors

Link to publication: https://rdcu.be/c7aDw

ABSTRACT

Objective: To investigate whether ensembles that combine prognostic models developed using different databases (a simple distributed learning approach) perform better in new data than single database models?

Materials and Methods: For a given prediction question we trained five single database models each using a different observational healthcare database. We then developed and investigated numerous ensemble models that combined the different database models. Performance of each model was investigated via discrimination and calibration using a leave one dataset out technique, i.e., hold out one database to use for validation and use the remaining four datasets for model development. The internal validation of a model developed using the hold out database was calculated and presented as the 'hypothetical optimum' for comparison.

Results: Fusion ensembles generally outperformed the single database models and were more consistent when applied to new data. Stacking ensembles performed poorly in terms of discrimination when the labels in the new data were limited. Calibration was poor when ensembles and single database models were applied to new databases.

Discussion: Fusion ensembles appear to improve model performance in new data and there was little difference in discrimination performance across fusion frameworks. Therefore, the simple weighted fusion may be preferable. Differences may occur when more database models are combined if some of the models perform poorly. Stacking ensembles may improve calibration but require a sufficient number of labels in the new data, which is a limitation.

Conclusion: A simple distributed learning approach using ensembles that combine models developed independently across databases for the same prediction question may improve the discriminative performance in new data but will need to be recalibrated.

BACKGROUND AND SIGNIFICANCE

Big observational healthcare databases, such as insurance claims data or electronic healthcare records, often contain data on large and diverse populations. One area where these datasets may benefit healthcare is in the application of machine learning to develop prognostic models. Prognostic models aim to predict a patient's risk of experiencing some future event (e.g., cardiovascular illnesses) [1] based on their current and historic health. In general, a prognostic task can be decomposed into three parts, the target population/index, the outcome, and the time-at-risk [2]. The target population is the set of patients for whom you attempt to predict the risk of some future outcome and the index is the point in time you want to make the prediction. The outcome is the medical event you want to predict, and the time-at-risk is the time interval (relative to the index) you want to predict the outcome occurring within. Prognostic models are learned from observational healthcare databases by finding patients in the database who historically match the target population, determining features such as age, gender, and medical history at index for each patient and then observing whether they had the outcome during the time-at-risk. Supervised learning, such as binary classification, is then applied to learn the differences between the people who had the outcome during the time-at-risk vs the people who did not. Often the aim is to develop a model using the historical data but apply the model to current patients to calculate a probability of whether they will have the outcome during the future time-at-risk. Such models could improve healthcare by informing medical decision making, but only if these models perform sufficiently well when implemented in their intended setting. For example, a model intended to be used by a family medicine doctor to help them decide which patients should be given preventative medicine may be developed using a large insurance claims database but needs to transport well into the family medicine setting. The performance in a new database (transportability) of a model is initially assessed by externally validating a model across diverse datasets with different patient case mixes [2,3]. It is common for a model's performance to deteriorate when transported to a different database [2]. The deterioration in performance may be due to the model or the differences between the development and validation populations [4]. A model that transports well to other databases is much more valuable in clinical practice. The question is how to best develop models with high transportability?

Big observational healthcare datasets only contain a sample of the population. This is frequently a non-random sample, for example the data may over sample (or only contain) certain ethnicities, genders, ages or patients with low/medium/high wealth. If a database used to develop a prognostic model contains a non-random sample of the target population then this will most likely negatively affect its performance if applied on the full population. However, different datasets, with varying patient case mixes, may give diverse perspectives when developing prognostic model for the same prediction task. Learning models across different healthcare datasets (e.g., a US insurance claims database, a UK primary care database and a US electronic healthcare record database), known as distributed learning [5-8], may lead to more transportable models.

There are numerous approaches to distributed learning: i) combine the datasets and develop a model using the combined data, ii) apply a distributed algorithm that iterates across datasets or iii) combine models developed using different datasets. The first option is generally limited as sharing patient-level data between researchers is often not possible due to privacy restrictions and therefore it is not possible to train a single model using the combination of different datasets. The second option is limited due to the administration required if the algorithm needs to communicate with each dataset (held at different physical locations by different owners) multiple times. Although some distributed algorithms that only require access to each database once, termed 'one-shot distributed algorithms', exist for certain generalized linear models [8]. The one-shot approach is not suitable for most machine learning models. The third option is most feasible, as it is possible for researchers to easily share prognostic models, they develop using their own data and these models could be combined via ensemble techniques (ensemble modelling is the common machine learning approach used to combine binary classification models). This prompts the question; can we implement the third option and combine models developed using diverse datasets to improve model transportability in new data (e.g., in a clinical setting)?

Ensemble learning is the process of producing multiple models, potentially pruning the set of models, and then combining the remaining models [9]. Often the ensemble increases both model accuracy and performance stability compared to any single classifier [10]. Ensembles either combine homogeneous models (same learning algorithm) or heterogeneous models (different learning algorithms). Homogeneous ensembles use the same learning algorithm but modify the perspective by using different training data (e.g., different instances, different features or by adding noise), different metrics or using different model settings (e.g., hyper-parameter values). Heterogeneous models take a different perspective as each learning algorithm makes different assumptions about the data. Combining the models is often done by fusing the classifiers [11], stacking [12] or using a mixture of experts [13]. Examples of simple fusing classifiers include i) majority vote, the combination technique used by random forest 'bagging' [14], ii) calculating the mean prediction value across classifiers or iii) weighted mean of the classifier predictions based on performance measures. Weighing each classifier's prediction based on performance is better than taking the mean of all classifier predictions when the classifier performances differ (e.g., one classifier is better than the others) [11]. A mixture of experts is similar to weighted mean but instead of using universal weights across the instances, the weights are assigned per instance [13]. These ensembles are considered independent ensemble frameworks, as the models are trained independently and then combined [10]. A more advanced independent ensemble framework is known as 'stacking'. Stacking is a meta-combination method that uses the set of models' predictions as features and trains a new model that learns to predict the outcome using these prediction features [12]. A limitation of stacking is that it requires additional labelled data to learn how to best combine the individual models. Alternatively, 'dependent ensemble' frameworks train classifiers sequentially and each classifier depends on the output of the prior classifier [10]. Boosting is a dependent fusion ensemble framework as models are sequentially trained, and weights are assigned to the objective function of each model during training based on prior models' mistakes [15]. The above examples are just a selection of the commonly used combination methods and there are numerous other ways to combine the models [10].

Objective

This paper aims to determine whether prognostic model ensembles that combine regularized logistic regression models independently developed across different healthcare databases perform better in new data (more transportable) than each individual database prognostic model (single dataset model). A model with improved transportability is likely to also perform better when used clinically for decision making.

MATERIALS AND METHODS

The Observational Health Data Science and Informatics (OHDSI) PatientLevelPrediction framework is used throughout this paper [2] for developing prognostic models using observational healthcare data.

Databases

Four US claims and an EHR databases are explored, see Table 1.

Table 1 Summary of the five databases used in this study

Name	Type	Description	Start	End	Size (million lives)
IBM Medicare Supplemental Beneficiaries (MDCR)	US Claims	Patients aged 65 or older with supplemental healthcare.	2000-01-01	2019-12-31	10.115
IBM Medicaid (MDCD)–	US Claims	Patients with government subsidized healthcare.	2006-01-01	2018-12-31	28.777
Optum® De-Identified Clinformatics® Data Mart Database (Optum Claims)	US Claims	Patients of all ages	2000-05-01	2019-12-31	84.310
IBM Commercial Claims and Encounters (CCAE)	US Claims	The patients in this database are aged 65 or younger. They are employees who receive health insurance through their employer and their dependents.	2000-01-01	2019-12-31	152.96
Optum® de-identified Electronic Health Record Dataset (Optum EHR)	US EHR	Patients of all ages	2006-01-01	2019-03-31	96.505

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

All datasets used in this paper were mapped into the OHDSI Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) version 5 [16]. The OMOP-CDM was developed to enable researchers with diverse datasets to have a standard database structure. This enables analysis code and software to be shared among researchers which facilitates external validation of prediction models.

Prediction problem

As an example, the problem: "Amongst patients with pharmaceutically-treated depression, which patients will develop <an outcome> during the 1-year time interval following the start of the depression episode?" is investigated.

The target population of pharmaceutically treated depressed patients is defined as: patients with a condition record of major depressive disorder and the index date was the first record date. Inclusion criteria are:

- Antidepressant recorded within 30 days before to 30 days after the target population index date
- No history of psychosis
- No history of dementia
- No history of mania
- >=365 days prior observation

Twenty-one models predicting 21 different outcomes occurring for the first time between 1 day after index until 1 year after index are developed. The 21 outcomes are: acute liver injury, acute myocardial infarction, alopecia, constipation, decreased libido, delirium, diarrhoea, fracture, gastrointestinal hemorrhage, hyponatremia, hypotension, hypothyroidism, insomnia, nausea, seizure, stroke, sudden cardiac death, suicide and suicidal ideation, tinnitus, ventricular arrhythmia and vertigo.

The above definition of prediction problem is the same as used in reference [2].

In this study a random sample of 500,000 patients from the target population (>1 million patients in Optum claims, >2 million patients in Optum EHR and >2 million in CCAE) were used throughout the study. This improved the efficiency of model development and also resulted in some low outcome counts, enabling the investigation into whether the outcome count impacts the ensemble performance.

Labelled data

We constructed labelled datasets for each database and outcome pair. For the *i*th patient in database *k* we used one-hot-encoding to create binary features indicating the presence of any medical condition or drug recorded prior to index (first record of major depressive disorder) and extracted the patient's gender and age at index (in 5-year bins). Let $\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{k}}$ represent the feature vector for the *i*th patient in database *k*. Labels were determined per outcome, with y_{ij}^{k} corresponding to the presence $(y_{ij}^{k} = 1)$ or absence $(y_{ij}^{k} = 0)$ of outcome *j* in the year after index for patient *i* in database *k*. This resulted in 105 labelled datasets $\{(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{k}}, y_{ij}^{k})\}_{i}$.

Statistical analysis

Binary Classifiers (Level 1 models)

For each database and outcome, a regularized logistic regression model with least absolute shrinkage and selection operator (LASSO) penalization was trained [17] using 80% of the data to develop the model and 20% of the data were held out to internally estimate the model performance (test set performance). Three-fold cross validation was applied in the 80% development data to learn the optimal regularization value. The final LASSO logistic regression coefficients were learned with the optimal hyper-parameter set using all of the 80% development data.

Let $f_{ij}(\boldsymbol{x}): R_m \rightarrow [0,1]$ correspond to the Level 1 logistic regression model that was developed using the *i*th database (database *i*) to predict the *j*th outcome (outcome *j*), where **x** is the m-dimension feature vector for a patient. Given a patient's feature vector, the Level 1 model developed in database *i* predicts a value between 0 and 1 that corresponds to the probability that the patient will experience outcome *j*.

Performance Evaluation

Internal validation is when a model is developed and evaluated in the same database and external validation is when a model is developed and evaluated in different databases. For both internal and external validation, model discrimination and calibration were calculated. Model discrimination assesses how well a model ranks patients based on risk, this was calculated using the area under the receiver operating curve (AUROC). The AUROC is a ranking measure that corresponds to the probability that if a non-outcome patient was sampled and an outcome patient was sampled, the predicted risk assigned to the outcome patient is greater than the predicted risk assigned to the non-outcome patient. An AUROC of 0.5 corresponds to randomly predicting risk (no discriminative ability) and an AUROC of 1 corresponds to perfect prediction (a higher risk is predicted for all patients who will experience the outcome compared to those who will not). Calibration assesses how closely the predicted risk matches the true risk. For example, if a model is well calibrated, then if 10 patients are assigned a 10% risk, only 1 of them should experience the outcome. In this study, calibration was calculated using calibration-in-the-large [18] which compares the model's mean predicted risk in the population with the observed risk (a model is considered well calibrated if the mean predicted risk matches the observed risk in the population).

The internal validation of each Level 1 model (test set performance) provides a benchmark performance for the database and outcome pair. The internal validation of each Level 1 model, trained in database *k* to predict outcome *j*, was determined by calculating the AUROC and calibration-in-the-large using the predicted risk $f_{ij}(\boldsymbol{x}_{\boldsymbol{i}}^{\boldsymbol{k}})$ and the true label y_{ij}^{k} for each patient in the 20% held out set (test set).

Binary Ensemble Classifiers (Level 2 models)

The ensembles in this study combine the Level 1 models developed in the different databases that predict the same outcome. Generally, an ensemble that predicts outcome *j* is a function of the *N* Level 1 models that predict outcome *j*:

$$f_j(\boldsymbol{x}) = g(\{f_{ij}(\boldsymbol{x})\}_{i\in\{1,2,\ldots,N\}})$$

Seven different ensemble approaches were investigated to combine the Level 1 models, that predict the same outcome (*j*) but are trained on *N* different databases $(\{f_{kj}\}_{k\in\{1,2,\ldots,N\}})$, using different heuristics.

A weighted fusion ensemble to predict the outcome *j* combines the Level 1 models by assigning each Level 1 model a weight:

$$f_j(\boldsymbol{x}) = \sum_{i} w_{ij} f_{ij}(\boldsymbol{x})$$

where w_{ij} is the weight assigned to the Level 1 model trained using database *i* to predict outcome *j*. In this study different weighting heuristics are investigated:

1. Mean Ensemble (**mean**) – for a patient, their predicted risk is the mean of the predicted risks of the included Level 1 classifiers (equal weighting so w_{ij} = 1/N, where *N* is the number of models being combined)
2. AUROC Ensemble normalized weights (**auc1**) – for a patient, their predicted risk is a weighted mean of the predicted risks of the included Level 1 models, where each Level 1 model's weight is based on the model's internal area under the receiver operating characteristic curve (AUROC) that was calculated in the 20% held out data.The weights are scaled relative to an AUROC of 0.5 and normalized to ensure the total weight across models was 1 (AUROC performance weighting so $w_{ij} = \frac{|AUROC_{ij}-0.5|}{\sum_k |AUROC_{kj}-0.5|}$), where AUROC$_{ij}$ is the internal AUROC value for the Level 1 model developed in database *i* to predict outcome *j*.
3. AUROC Ensemble unnormalized weights (**auc2**) – similar to 2) a patient's risk is a weighted mean of the predicted risks of the included Level 1 models, where each Level 1 model's weight is based on the model's internal AUROC. The weights are scaled between 1 for perfect discrimination and -1 for models that predict the opposite labels perfectly ($w_{ij} = \frac{AUROC_{ij}-0.5}{0.5}$), where AUROC$_{ij}$ is the internal AUROC value for the Level 1 model developed in database *i* to predict outcome *j*.
4. Similarity Weighted Ensemble (**sim**)- for a patient, their predicted risk is a weighted mean of the predicted risks of the included Level 1 models, but weights are based on how similar the Level 1 model's development population mean value for each predictor are compared to the population that the patient is in. The cosine similarity metric was used for the two

vectors containing the mean values in the patient's dataset and the Level 1 model's development data (case mix similarity weighting $w_{ij} = \frac{cosine(\boldsymbol{d},\boldsymbol{d_i})}{\sum_k cosine(\boldsymbol{d},\boldsymbol{d_k})}$) where **d** is an m-dimensional vector corresponding to the mean values of the features included in model f_{ij} in the database the ensemble is being applied to and $\mathbf{d_i}$ is an m-dimensional vector corresponding to the mean values of the features included in model f_{ij} in database *i*.

5. Age Weighted Ensemble (**age**)– for a patient, their predicted risk is a weighted mean based on how similar the model development data population mean age was compared to the patient's population mean age (case age similarity weighting $w_{ij} = \frac{d(age,age_i)}{\sum_k d(age,age_k)}$), where *age* is the mean age in years of the patients in the dataset the model is being applied to, age_i is the mean age of the patients in database *i* and $d(age,age_i) = 1/(1+|age - age_i|)$.

The mixture of expert ensembles $f_j(\boldsymbol{x})$ use the equation: $f_j(\boldsymbol{x}) = \Sigma_i g_{ij}(\boldsymbol{x}) f_{ij}(\boldsymbol{x})$ where g_{ij} is the gating function value for Level 1 model developed in database *i* to predicted outcome *j*.

6. Age Mixture of Experts Ensemble (**ageME**) – for a patient, their predicted risk is calculated using the Level 1 model developed using a population with a mean age that most closely matches the patient's age, the gating function is:

$$g_{ij}(\boldsymbol{x}) = \begin{cases} \mathbf{1}, \ \boldsymbol{i} \equiv \min\limits_{\boldsymbol{k}}(\boldsymbol{age_k} - \boldsymbol{age}) \\ \mathbf{0}, \qquad \boldsymbol{otherwise} \end{cases}$$

Where age_k is the mean age in years of the patients in database *k* and *age* is the age in years of the patient whose risk is being calculated.

Stacking ensembles involved training a Level 2 model that uses the Level 1 model predictions as features.

7. Stacking ensemble –a LASSO logistic regression model was trained as the Level 2 model that used the predicted risk from each Level 1model as predictors (effectively this learned the Level 1 model weightings).The stacking ensemble requires labelled data in the validation dataset whereas the other ensembles do not require this. As it is often not possible to get large amounts of labelled data in the validation dataset or application dataset, it was investigated how well the stacking ensemble would do if i) only 1,000 patients (**s|1000**), ii) only 10,000 patients (**s|10000**) and iii) all available patients (**s|All**) were used to learn the weightings.

Model Transportability

For each ensemble model a leave-one-database out approach was used to estimate external validation when the ensemble was transported to new data. Figure 1 illustrates the leave-one-database out approach. For example, to estimate the mean fusion ensemble performance in predicting insomnia when externally validated on MDCR, the Level 1 models trained on the MDCD, CCAE, Optum Claims and Optum EHR to predict insomnia were applied to each patient in MDCR and then the mean of the patient's predicted risks across the four Level 1 models

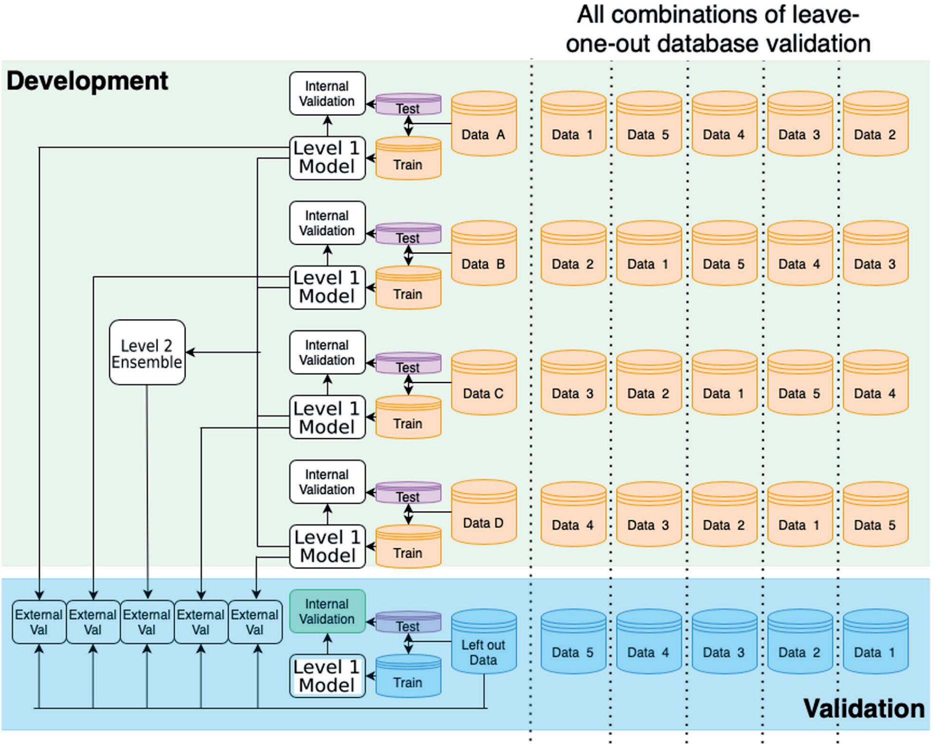


Figure 1 The leave-one-database-out design used to evaluate the transportability of the Level 1 models trained using a single database and the Level 2 ensembles that combine multiple Level 1 models.

The figure shows that five different combinations were used, where four of the five databases were used to develop the models and the final database was used to fairly evaluate the transportability of the models. In addition, a model was trained using the left-out database to calculate the internal validation that could be considered the 'hypothetical optimum' performance for the database given sufficient training data. We compared how similar the external validation of each model was with the 'hypothetical optimum' benchmark.

was calculated per patient. The mean fusion ensemble predictions are then validated using the ground truth labels in the left-out database where it was known which patients experienced insomnia. This was repeated five times by leaving each database out once.

Denoting the set of feature and label pairs in database *k* for outcome *j* as: $\{(\boldsymbol{x_i^k}, y_{ij}^k)\}_i$, the vector of predicted risks in database *k* for outcome *j* using the mean fusion ensemble but excluding the database *k* model is:

$$\boldsymbol{pred_j^k} = (\sum_{i\neq k}\frac{f_{ij}(\boldsymbol{x_1^k})}{n}, \sum_{i\neq k}\frac{f_{ij}(\boldsymbol{x_2^k})}{n}, \dots, \sum_{i\neq k}\frac{f_{ij}(\boldsymbol{x_m^k})}{n})$$

The ground truth in database *k* is:

$$\boldsymbol{truth_j^k} = (y_{1j}^k, y_{2j}^k, \dots, y_{mj}^k)$$

The external AUROC and calibration metrics for the mean fusion ensemble applied to database *k* for outcome *j* is then calculated by comparing the predictions and ground truth labels.

$$externalAUROC_j^k = AUROC(\boldsymbol{pred_j^k}, \boldsymbol{truth_j^k})$$

In general, the predictions for all Level 1 models and Level 2 ensemble models when transported to database *k* are:

$$\boldsymbol{level1Pred_j^k} = \left(f_{ij}(\boldsymbol{x_1^k}), f_{ij}(\boldsymbol{x_2^k}), \dots, f_{ij}(\boldsymbol{x_m^k})\right), where\ i \neq k$$

$$\boldsymbol{ensemblePred_j^k} = (g\left(\{f_{ij}(\boldsymbol{x_1^k})\}_{i\neq k}\right), g\left(\{f_{ij}(\boldsymbol{x_2^k})\}_{i\neq k}\right), \dots, g\left(\{f_{ij}(\boldsymbol{x_m^k})\}_{i\neq k}\right))$$

To put the performance of the Level 1 and Level 2 models (that do not used database *k*) into context, the 'hypothetical optimal' performance that is achievable in database *k* was estimated. The 'hypothetical optimum' is defined as the internal validation performance (using a 20% test set $\{(\boldsymbol{\hat{x}_i^k}, \hat{y}_{ij}^k)\}_i)$ of the Level 1 model developed in database *k*:

$$\boldsymbol{internalPred_j^k} = (f_{kj}(\boldsymbol{\hat{x}_1^k}), f_{kj}(\boldsymbol{\hat{x}_2^k}), \dots, f_{kj}(\boldsymbol{\hat{x}_t^k}))$$

$$\boldsymbol{truthTest_j^k} = (\hat{y}_{1j}^k, \hat{y}_{2j}^k, \dots, \hat{y}_{tj}^k)$$

The internal AUROC in database *k* for outcome *j* is then:

$$internalAUROC_j^k = AUROC(\boldsymbol{internalPred_j^k}, \boldsymbol{truthTest_j^k})$$

Given sufficient data, the internal performance of a model can be considered the upper bound of achievable performance (conditional on the same features being available to internal and external model development). If a model transported to new data has an external performance close to the internal performance of a model developed using the data, then this can be considered to have transported well. Consequently, to determine how well a model transports the difference in performance between the internal validation AUROC of the Level 1 model trained using the left-out database, database *k*, and the external validation AUROC of models when applied to the left-out database *k* was calculated:

$$AUROC\_difference_j^k = externalAUROC_j^k - internalAUROC_j^k$$

where $externalAUROC_j^k$ is the performance of the model in database *k* (trained without dataset *k*) in predicting outcome *j* and $internalAUROC_j^k$ is the Level 1 model predicting outcome *j* trained in database *k's* performance on the 20% test set. To show how well each model transports in general, box plots were created to show the distribution of $AUROC\_difference_j^k$ across the different outcomes and databases. Distributions centered around 0 indicate excellent transportability and distributions with a small range indicate consistency.

RESULTS

The data sizes are presented in Table 2 and the database characteristics are displayed in Table 3. The smallest target population was the one extracted from the MDCR database, and this population were older and had higher rates of cancer and cardiovascular issues prior to index. The MDCD target population was the youngest and also had the highest rate of obesity recorded in the prior year. In general, the characteristics varied greatly across the datasets, indicating different patient case-mixes. The outcome count was generally greater than 100 except for delirium in MDCD and Optum Claims and Seizure in MDCR and Optum EHR.

Figure 2 presents box plots of the *AUROC_differences* per Level 1 model (non-ensemble) and Level 2 model (ensemble) when transported to each held out database across the 21 outcomes. The non-ensemble box plots show a lower median value and greater range of values compared to the fusion ensembles. The fusion ensembles achieved discriminative performances similar to

Table 2 - The outcome counts and percentage of target population who develop the outcome during the tine-at-risk

Outcome	**CCAE (N ~499,678) (%)**	**MDCR (N ~160,956) (%)**	**MDCD (N ~469,302) (%)**	**Optum EHR (N ~499,881) (%)**	**Optum Claims (N ~499,753) (%)**
Acute liver injury	14875 (3.35)	7226 (5.4)	21654 (5.47)	18535 (4.18)	18619 (4.31)
Acute myocardial infarction	1494 (0.3)	935 (0.59)	3800 (0.83)	816 (0.16)	1298 (0.26)
Alopecia	10672 (2.32)	7569 (5.64)	20597 (5.2)	16597 (3.69)	16571 (3.75)
Constipation	4170 (0.85)	6399 (4.39)	9210 (2.05)	10192 (2.13)	10282 (2.16)
Decreased libido	491 (0.1)	1080 (0.69)	905 (0.19)	287 (0.06)	708 (0.14)
Delirium	174 (0.03)	510 (0.32)	86 (0.02)	267 (0.05)	91 (0.02)
Diarrhoea	1661 (0.34)	130 (0.08)	785 (0.17)	1210 (0.24)	1603 (0.32)
Fracture	509 (0.1)	963 (0.61)	894 (0.19)	381 (0.08)	758 (0.15)
Gastrointestinal haemorrhage	985 (0.2)	1298 (0.81)	1666 (0.36)	356 (0.07)	1021 (0.2)
Hyponatremia	19754 (4.65)	7824 (5.95)	33518 (9.82)	24043 (5.65)	23304 (5.67)
Hypotension	380 (0.08)	1153 (0.74)	636 (0.14)	230 (0.05)	683 (0.14)
Hypothyroidism	297 (0.06)	642 (0.4)	1056 (0.23)	162 (0.03)	333 (0.07)
Insomnia	3046 (0.62)	2086 (1.38)	2468 (0.53)	3049 (0.62)	4114 (0.85)
Ischemic stroke all inpatient	3120 (0.64)	1824 (1.19)	2655 (0.57)	2775 (0.56)	4139 (0.85)
Nausea	2722 (0.56)	4071 (2.77)	4033 (0.89)	4368 (0.9)	5846 (1.22)
Open angle glaucoma	6117 (1.33)	3853 (2.83)	5374 (1.22)	8786 (2.03)	9943 (2.33)
Seizure	184 (0.04)	67 (0.04)	307 (0.07)	94 (0.02)	199 (0.04)
Suicide and suicidal ideation	10221 (2.13)	993 (0.62)	21518 (5.09)	9957 (2.1)	8063 (1.67)
Tinnitus	2628 (0.53)	4276 (2.87)	5082 (1.12)	6920 (1.44)	7643 (1.62)
Ventricular arrhythmia and sudden cardiac death	20806 (4.91)	6846 (5.12)	27233 (6.92)	23655 (5.6)	23772 (5.89)
Vertigo	2577 (0.53)	748 (0.47)	2269 (0.49)	2341 (0.48)	2782 (0.57)

Table 3 - characteristics of the target population (patients with depression initiating treatment) per database

	CCAE	MDCD	MDCR	Optum Claims	Optum EHR
Mean Age	41	35	75	50	49
Male (%)	30.8	25.9	32.2	31.7	29.2
Mean number outpatient visits in prior year	16.3	31.2	26.8	16.6	32.4
Frequency of patients experiencing condition in prior year:					
Pain	0.60	0.74	0.74	0.66	0.57
Anxiety	0.41	0.50	0.28	0.42	0.43
Acute inflammatory disease	0.32	0.36	0.24	0.31	0.18
Neoplastic disease	0.22	0.14	0.46	0.27	0.17
Essential hypertension	0.25	0.31	0.69	0.40	0.37
Obesity	0.11	0.19	0.11	0.13	0.17
Heart disease	0.09	0.14	0.46	0.20	0.18
Diabetes mellitus	0.09	0.14	0.27	0.16	0.16
Urinary tract infectious disease	0.09	0.14	0.16	0.12	0.07
Anemia	0.07	0.12	0.20	0.12	0.11

the 'hypothetical optimum' when transported to new databases (*AUROC_difference* values close to 0). The age-based mixture of expert and stacking ensembles that used 1,000 or 10,000 labels generally performed worse than the non-ensembles in terms of discrimination when transported. The stacking ensemble using all the labelled data available achieved external AUROC similar to the 'hypothetical optimum' but was not better than the fusion ensembles. The full external validation discrimination performance across the 21 outcomes and 5 databases for the non-ensembles and ensembles are presented in online Appendix A.

Each calibration in the large (the mean predicted risk) is presented in Figure 3 and Figure 4. The calibration in the large plots show the mean predicted risk per Level 2 model (ensemble) or Level 1 model (non-ensemble) and the dashed horizontal line is the observed population risk. A model is well calibrated if the mean predicted risk matches the observed population risk. Figures 3-4 show that the mean predicted risks did not often match the observed population risk, except for the stacking ensemble.

DISCUSSION

The results show that weighted fusion ensembles that combine multiple prognostic models developed in different databases appear to have more stable discriminative performances when transported to new databases compared to the Level 1 (single database) models. However, calibration appears to be an issue for all models that are transported to new databases (except stacking ensembles with sufficient labels).

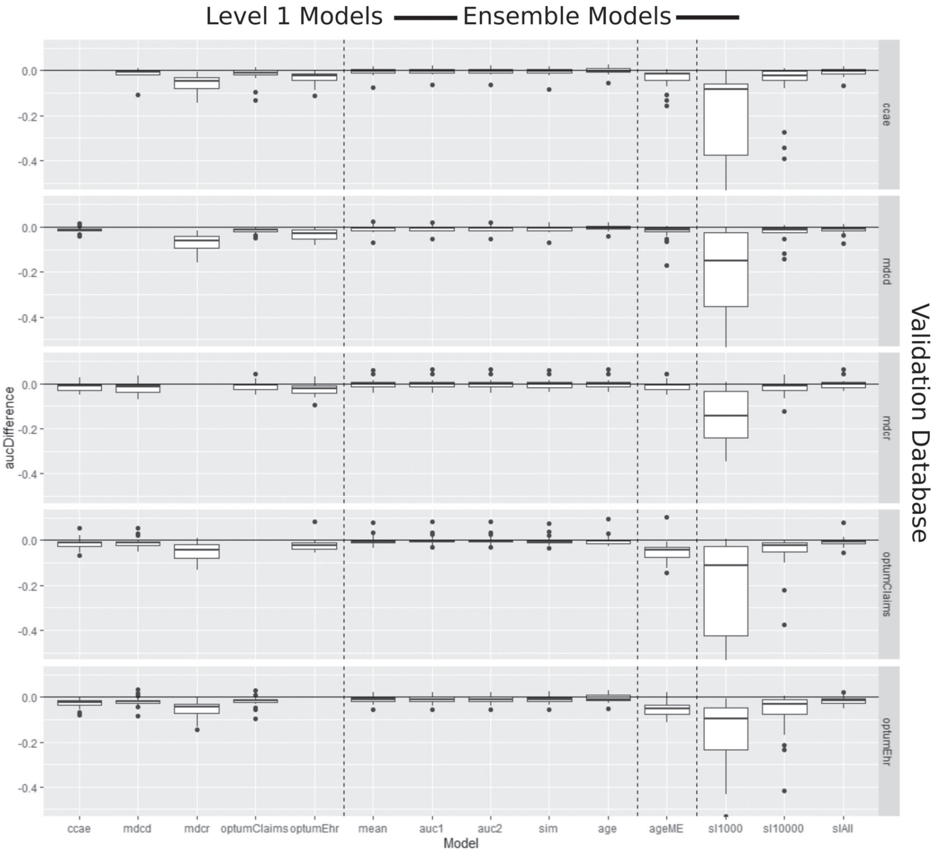


Figure 2 Box plots showing the difference between the external validation AUROC minus the internal validation AUROC per non-ensemble (Level 1 model) and ensemble method (Level 2 model) across the five databases. The rows represent the external database (the database that was excluded from the model/ensemble development) that was used to fairly evaluate the models/ensembles. The x-axis represents the model/ensemble technique. Box plots centred around 0 with a small range indicate highly transportable and consistent external discriminative performance. The dashed vertical lines separate the non-ensembles, the fusion ensembles, the mixture of expert ensembles and the stacking ensembles.

This study showed that certain ensembles combining models developed independently across difference databases transport better than the Level 1 single database models. The weighted fusion ensembles and stacking ensemble (that used all data) consistently achieved discrimination close to the 'hypothetical optimum' in the new data whereas the Level 1 single models generally performed slightly worse than the 'hypothetical optimum'. The Level 1 single database models were also less consistent across outcomes and certain database models did better than others (e.g., Optum claims models transported better than MDCR models). This variability may be due to each database containing diverse patient case-mixes, as seen in Table 3. The ensembles can combine the perspectives of the Level 1 models trained with different populations making them more robust to new populations. The calibrations of the transported

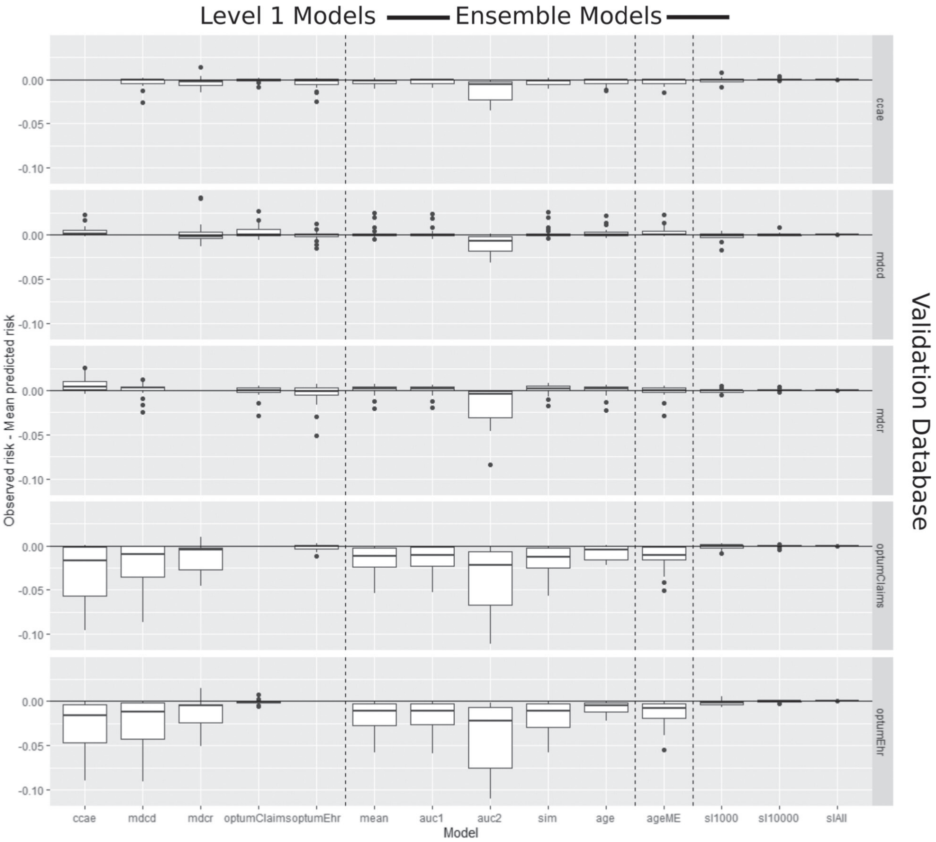


Figure 3 - The calibration-in-the-large (mean predicted risk) for each non-ensemble (Level 1 model) and ensemble (Level 2 model) when externally validated. The rows represent the prediction problem (different outcomes) and the columns represent the external validation databases. The x-axis represents the different models/ensemble techniques. The solid horizontal line is the observed population risk in the external database. The dashed vertical lines divide the non-ensembles, the weighted ensembles, the mixture of expert ensemble and the stacker ensembles. A model is well calibrated when externally transported if the top of the bar is near to the solid horizontal line.

models were generally poor, except the stacking model (using all data) as this used labelled data so was effectively recalibrated. If all the Level 1 single database models are mis-calibrated, then it makes sense that any ensemble combining them would also be mis-calibrated. This highlights the importance of model recalibrating before implementing them in new patient populations. It may be possible to recalibrate without labelled data by changing the intercept based on how common the outcome is in the target population the model is being applied to. If labels are available for some patients, then standard recalibration techniques can be implemented.

The results show the type of ensemble heuristic impacted transportability. The ensembles that performed the best in terms of discrimination when transported were the weighted fusion ensembles. The stacking ensemble did almost as well as the weighted fusion ensemble when there were sufficient labels, but it required labels in the new data it is being transported to

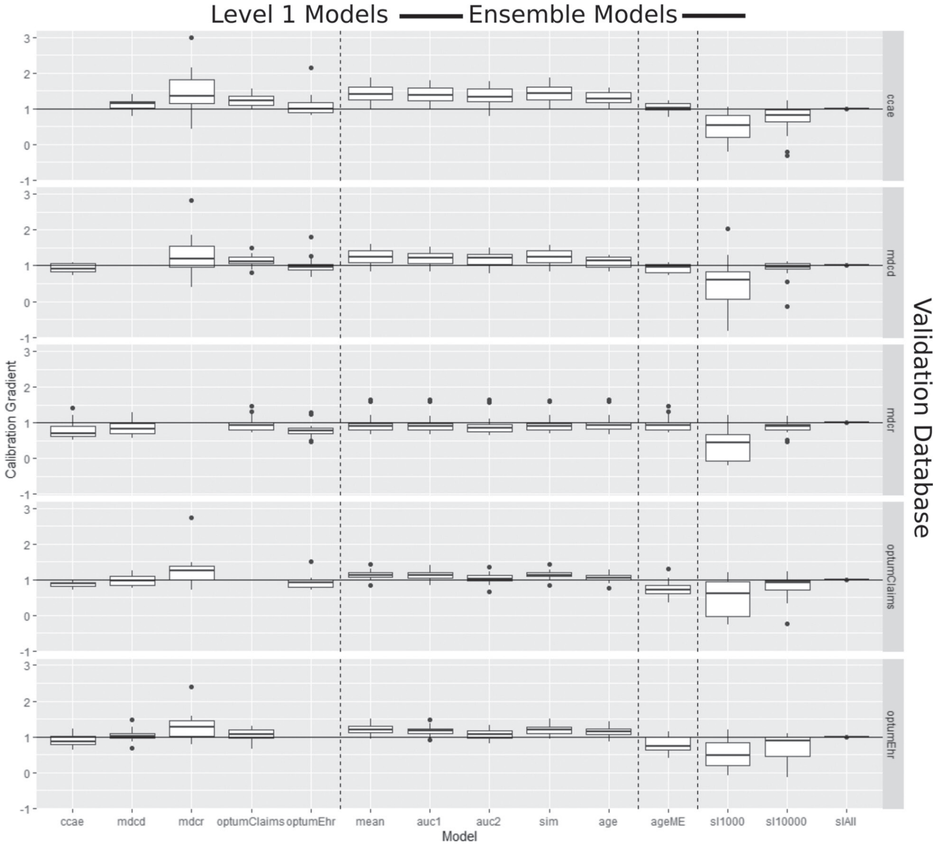


Figure 4. The calibration-in-the-large (mean predicted risk) for each non-ensemble (Level 1 model) and ensemble (Level 2 model) when externally validated. The solid horizontal line is the observed population risk in the external database. The dashed vertical lines divide the non-ensembles, the weighted ensembles, the mixture of expert ensemble and the stacker ensembles.

whereas the weighted fusion ensembles did not. Requiring labels is a big disadvantage and therefore the weighted fusion ensembles are more useful. Interestingly, the simple mean fusion ensemble (uniform weighting) was comparable to the AUROC, age and database similarity weighted ensembles. Due to its simplicity, the mean fusion ensemble shows promise at being able to lead to more transportable prognostic models. If it worth noting, the age weighted ensembles may have benefitted in this study by the databases being similar (mostly US claims databases). For example, Optum claims is a mixture of patients that are similar to the patients in CCAE and MDCR, Therefore the age weighting may not perform well when the databases are more diverse. The weighted fusion ensembles and mixture of expert ensemble may have been impacted by the outcome rate differing between the databases. If the outcome is more common in a database, then a logistic regression model's intercept is likely to be greater and the model's mean predicted risk is likely to be higher than a model trained in data with fewer outcomes. This effectively may add more weighting to Level 1 models trained in databases that have a higher outcome percentage in the data.

The key advantage of this study is that we were able to compare the transportability of Level 1 models (developed in a single database) and ensembles combining Level 1 models developed in different databases across many prediction problems and across five datasets. In total we trained 21 (outcomes) x 5 (databases) single database models and created 21 (outcomes) x 5 (databases) x 7 (ensemble methods) ensemble models. The limitation of this study is the generalizability of findings as we only investigated one target population and we only used US data. In future work it would be useful to repeat this experiment across different target populations and externally validate the models (ensemble/non-ensemble) developed in this study across non-US databases. The OHDSI network and collaboration could be used to scale up this study across more diverse databases in future work [19]. In addition, there are numerous ways to combine the Level 1 models into an ensemble and we only investigated 7 simple approaches. However, these results provide a benchmark for comparing other ensembles techniques.

In this study 500,000 patients were sampled from each database as this provided a range of outcome sizes for the 21 outcomes investigated and enabled us to investigate the impact of outcome count in the study. Predicting rare outcomes is often an area of interest in healthcare and this may be where learning across multiple databases is more advantageous.

In future work it would be interesting to investigate whether rescaling the Level 1 models' predictions within the ensemble, to make the mean predicted risk for each Level 1 model within the ensemble equal, could improve the weighted fusion or mixture of expert ensembles. In addition, in this study we did not investigate pruning the Level 1 models within the ensembles, but this is an area of future research that may further improve transportability of an ensemble. In this study none of the Level 1 single database models achieved an AUROC ~0.5, but it may make sense to prune such models if the situation arises.

CONCLUSION

In this study we performed a large-scale empirical evaluation to investigate the transportability of a simple and feasible distributed learning approach that combines models developed in different databases via simple ensemble techniques. The results show that a mean fusion ensemble appears to transport to new data with higher discrimination compared to models developed in any single database. As a consequence, developing a mean fusion ensemble of prognostic models developed using different databases may lead to more clinically robust and useful prognostic models. However, recalibration is likely to be required.

Supplementary Information

The online version contains supplementary material available at
https://doi.org/10.1186/s12911-022-01879-6.

Authors' contributions

JMR, RDW and MJS contributed to the conception and design of the study. JMR and PBR contributed to the acquisition of data. All authors contributed to the analysis and interpretation of data. All authors contributed to drafting the article and revising it critically for important intellectual content. All authors contributed to the final approval of the version to be submitted.

Acknowledgments

None

Competing inerests

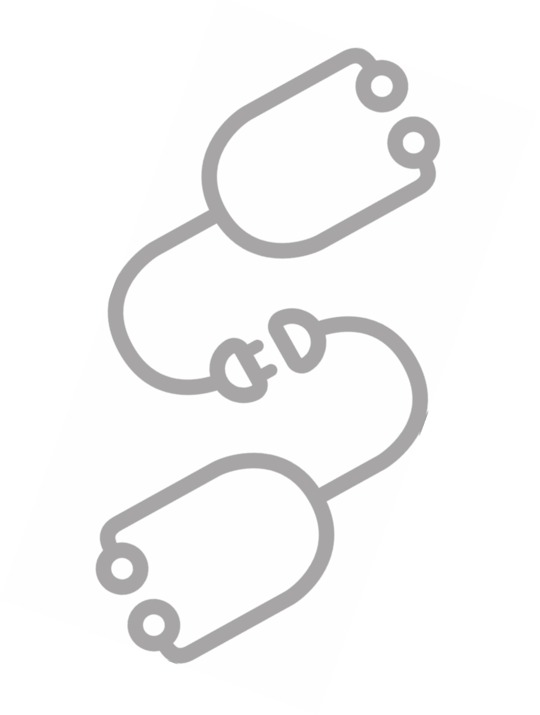
JMR, MJS, PBR are employees of Janssen Research and Development and shareholders of Johnson and Johnson.

Funding

PRR and RDW received funding from the European Health Data and Evidence Network (EHDEN) project of 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.

BIBLIOGRAPHY

1. Farzadfar, F. Cardiovascular disease risk prediction models: challenges and perspectives. Lancet Glob Health 2019;7(10):e1288-e1289.
2. Reps JM, Schuemie MJ, Suchard MA, et al. 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.
3. Debray TP, Vergouwe Y, Koffijberg H, et al. A new framework to enhance the interpretation of external validation studies of clinical prediction models. *J Clin Epidemiol* 2015;68(3):279-89.
4. Vergouwe, Y., Moons, K.G. and Steyerberg, E.W., External validity of risk models: use of benchmark values to disentangle a case-mix effect from incorrect coefficients. *Am J Epidemiol* 2010;*172*(8):971-980.
5. Jochems, A., Deist, T.M., Van Soest, J., Eble, M., Bulens, P., Coucke, P., Dries, W., Lambin, P. and Dekker, A., Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. *Radiotherapy and Oncology* 2016;*121*(3):459-467.
6. Bogowicz, M., Jochems, A., Deist, T.M., Tanadini-Lang, S., Huang, S.H., Chan, B., Waldron, J.N., Bratman, S., O'Sullivan, B., Riesterer, O. and Studer, G., Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer. *Scientific reports* 2020;*10*(1):1-10.
7. Deist, T.M., Dankers, F.J., Ojha, P., Marshall, M.S., Janssen, T., Faivre-Finn, C., Masciocchi, C., Valentini, V., Wang, J., Chen, J. and Zhang, Z., Distributed learning on 20 000+ lung cancer patients–The Personal Health Train. *Radiotherapy and Oncology* 2020;*144*:189-200.
8. Luo, C., Islam, M.N., Sheils, N.E., Reps, J.M., Buresh, J., Duan, R., Tong, J.M., Edmondson, M., Schuemie, M.J. and Chen, Y., 2020. Lossless Distributed Linear Mixed Model with Application to Integration of Heterogeneous Healthcare Data. *medRxiv*.
9. Tsoumakas, G., Partalas, I. and Vlahavas, I. A taxonomy and short review of ensemble selection. In *Workshop on Supervised and Unsupervised Ensemble Methods and Their Applications* 2008; 1-6.
10. Rokach, L. Ensemble-based classifiers. *Artif Intell Rev* 2010;**33**(1–2):1-39. doi:10.1007/s10462-009-9124-7.
11. Fumera, G. and Roli, F., Performance analysis and comparison of linear combiners for classifier fusion. In *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)* Springer, Berlin, Heidelberg. 2002:424-432.
12. Wolpert, D.H., Stacked generalization. *Neural networks* 1992;*5*(2):241-259.
13. Masoudnia, S. and Ebrahimpour, R. Mixture of experts: a literature survey. *Artif Intell Rev* 2014;*42*(2):275-293.
14. Breiman L. Random forests. *Mach Learn* 2001;45:5–32
15. Freund Y, Schapire RE. Experiments with a new boosting algorithm. Machine learning: proceedings of the thirteenth international conference 1996:325–332
16. Voss EA, Makadia R, Matcho A, et al. . Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *J Am Med Inform Assoc* 2015;223:553–564.
17. Suchard MA, Simpson SE, Zorych I, et al. . Massive parallelization of serial inference algorithms for complex generalized linear models. *ACM Transact Model Comput Simulation* 2013;231:10–32.
18. Van Calster B, Nieboer D, Vergouwe Y, et al. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clinl Epidemiol* 2016;74:167-76.
19. Hripcsak, G., Duke, J.D., Shah, N.H., et al. Observational Health Data Sciences and Informatics (OHDSI): opportunities for observational researchers. *Stud Health Technol Inform* 2015;*216*:574.



2

Seek COVER: Using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network

Ross D. Williams[1,†] MSc, Aniek F. Markus[1,†] MSc, Cynthia Yang[1] MSc, Talita Duarte-Salles[2] PhD, Scott L. DuVall [3] PhD, Thomas Falconer[4] MS, Jitendra Jonnagaddala[5] PhD, Chungsoo Kim[6] PharmD, Yeunsook Rho[7] PhD, Andrew E Williams[8] PhD, Amanda Alberga Machado[9], MPH, Min Ho An[10] MD, María Aragón[2] PhD, Carlos Areia[11] MSc, Edward Burn[2,12] PhD, Young Hwa Choi[13] MD PhD, Iannis Drakos[14] PhD, Maria Tereza Fernandes Abrahão[15] PhD, Sergio Fernández-Bertolín[2] MSc, George Hripcsak[4] MD, Benjamin Skov Kaas-Hansen[16,17] MD, Prasanna L Kandukuri[18] MS, Jan A. Kors PhD[1], Kristin Kostka[19] MPH, Siaw-Teng Liaw[5] MBBS PhD, Kristine E. Lynch, PhD[3], Gerardo Machnicki[20] PhD, Michael E. Matheny[21], MD , Daniel Morales[22] PhD, Fredrik Nyberg[23] MD PhD, Rae Woong Park[24] MD, PhD, Albert Prats-Uribe[12] MPH, Nicole Pratt[25] PhD, Gowtham Rao[26] PhD MD PhD, Christian G. Reich[19] MD PhD, Marcela Rivera[27] PhD, Tom Seinen[1] MSc, Azza Shoaibi[26] MPH PhD, Matthew E Spotnitz[4] MD, Ewout W. Steyerberg[28,29] PhD, Marc A. Suchard[30] MD PhD, Seng Chan You[24] MD, Lin Zhang[31,32] MD PhD, Lili Zhou[18] PhD, Patrick B. Ryan[26] PhD, Daniel Prieto-Alhambra[12] MD PhD, Jenna M. Reps[26,&] PhD, Peter R. Rijnbeek[1,&,*] PhD

†These authors contributed equally as co-first authors.
[&]These authors contributed equally as co-last authors.

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [2]Fundacio Institut Universitari per a la recerca a l'Atencio Primaria de Salut Jordi Gol i Gurina (IDIAPJGol); [3]Department of Veterans Affairs, University of Utah, Salt Lake City, UT, US,; [4]Department of Biomedical Informatics, Columbia University, New York, NY; [5]School of Public Health and Community Medicine, UNSW Sydney; [6]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea; [7]Department of Bigdata, Health Insurance Review & Assessment Service, Republic of Korea; [8]Tufts University School of Medicine, Institute for Clinical Research and Health Policy Studies, Boston, MA, 02111, USA US; [9]Independent Epidemiologist, OHDSI; [10]So Ahn Public Health Center, Wando County Health Center and Hospital, Wando, Republic of Korea; [11]Nuffield Department of Clinical Neurosciences, University of Oxford ; [12]Centre for Statistics in Medicine, NDORMS, University of Oxford; [13]Department of Infectious Diseases, Ajou University School of Medicine, Suwon, Republic of Korea; [14]Center for Surgical Science, Koege, Denmark; [15]Faculty of Medicine, University of Sao Paulo, Sao Paulo, Brazil; [16]Clinical Pharmacology Unit, Zealand University Hospital, Roskilde, Denmark ; [17]NNF Centre for Protein Research, University of Copenhagen, Denmark; [18]Abbvie, Chicago, United States; [19]Real World Solutions, IQVIA, Cambridge, MA, United States; [20]Janssen Latin America, Buenos Aires, Argentina; [21]Department of Veterans Affairs, USA; Vanderbilt University, USA; [22]Division of Population Health and Genomics, University of Dundee, UK; [23]School of Public Health and Community Medicine, Institute of Medicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg, Sweden; [24]Department of Biomedical Informatics, Ajou University School of Medicine, Suwon, Republic of Korea; [25]Quality Use of Medicines and Pharmacy Research Centre, University of South Australia, Adelaide, Australia; [26]Janssen Research & Development, Titusville, NJ, USA; [27]Bayer Pharmaceuticals, Bayer Hispania, S.L., Barcelona, Spain; [28]Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands; [29]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands [30]Department of Biostatistics, UCLA Fielding School of Public Health, University of California, Los Angeles, CA, USA; [31]School of Public Health, Peking Union Medical College, Beijing, China; [32]Melbourne School of Public Health, The University of Melbourne, Victoria, Australia.

Link to publication: https://rdcu.be/c7aDU

ABSTRACT

Background: We investigated whether we could use influenza data to develop prediction models for COVID-19 to increase the speed at which prediction models can reliably be developed and validated early in a pandemic. We developed COVID-19 Estimated Risk (COVER) scores that quantify a patient's risk of hospital admission with pneumonia (COVER-H), hospitalization with pneumonia requiring intensive services or death (COVER-I), or fatality (COVER-F) in the 30-days following COVID-19 diagnosis using historical data from patients with influenza or flu-like symptoms and tested this in COVID-19 patients.

Methods: We analyzed a federated network of electronic medical records and administrative claims data from 14 data sources and 6 countries containing data collected on or before 4/27/2020. We used a 2-step process to develop 3 scores using historical data from patients with influenza or flu-like symptoms any time prior to 2020. The first step was to create a data-driven model using LASSO regularized logistic regression, the covariates of which were used to develop aggregate covariates for the second step where the COVER scores were developed using a smaller set of features. These 3 COVER scores were then externally validated on patients with 1) influenza or flu-like symptoms and 2) confirmed or suspected COVID-19 diagnosis across 5 databases from South Korea, Spain, and the United States. Outcomes included i) hospitalization with pneumonia, ii) hospitalization with pneumonia requiring intensive services or death, and iii) death in the 30 days after index date.

Results: Overall, 44,507 COVID-19 patients were included for model validation. We identified 7 predictors (history of cancer, chronic obstructive pulmonary disease, diabetes, heart disease, hypertension, hyperlipidemia, kidney disease) which combined with age and sex discriminated which patients would experience any of our three outcomes. The models achieved good performance in influenza and COVID-19 cohorts. For COVID-19 the AUC ranges were, COVER-H: 0.69-0.81, COVER-I: 0.73-0.91, and COVER-F: 0.72-0.90. Calibration varied across the validations with some of the COVID-19 validations being less well calibrated than the influenza validations.

Conclusions: This research demonstrated the utility of using a proxy disease to develop a prediction model. The 3 COVER models with 9-predictors that were developed using influenza data perform well for COVID-19 patients for predicting hospitalization, intensive services, and fatality. The scores showed good discriminatory performance which transferred well to the COVID-19 population. There was some miscalibration in the COVID-19 validations, which is potentially due to the difference in symptom severity between the two diseases. A possible solution for this is to recalibrate the models in each location before use.

BACKGROUND

In early 2020 the growing number of infections due to the coronavirus disease 2019 (COVID-19) resulted in unprecedented pressure on healthcare systems worldwide and caused many casualties at a global scale. Although the majority of people had uncomplicated or mild illness (81%), some developed severe disease leading to hospitalization and oxygen support (15%) or fatality (4%)(1, 2). This presented a challenge both in finding effective treatments as well as in identifying which patients were at high risk and as such would benefit from protective measures. The most common diagnosis in severe COVID-19 patients was pneumonia, other known complications included acute respiratory distress syndrome (ARDS), sepsis, or acute kidney injury (AKI)(1).

The WHO Risk Communication Guidance distinguished two categories of patients at high risk of severe disease: those older than 60 years and those with "underlying medical conditions", which is non-specific(3). Using general criteria to assess the risk of poor outcomes is a crude risk discrimination mechanism as entire patient groupings are treated homogeneously ignoring individual differences. Prediction models can quantify a patient's individual risk and data-driven methods could help to identify risk factors that have been previously overlooked. However, a systematic review evaluating all available prediction models for COVID-19(4) concluded that despite the large number of prediction models being developed for COVID-19, none were considered ready for clinical practice. These COVID-19 prediction models were criticized for i) being developed using small data samples, ii) lacking external validation, and iii) being poorly reported.

In this article, we describe a process of using a proxy disease to develop a prediction model for another disease. This can be used in situations where there is a data scarcity for the disease of interest. In this process a model is developed using big data from a proxy disease and then assessed in the target disease. This preserves all the target disease data for validation to provide a more robust and reliable assessment of model performance in the intended setting. This increases the evidence of the performance of a model in the target disease compared to if the same data had been used for development. We describe a use-case for this process using influenza data to develop a model in the early stages of the COVID-19 pandemic. It has been well documented that influenza and COVID-19 have significant differences(5, 6). However, we aim to show that influenza data can be used to develop a well performing model that could have been transported and used in early COVID-19 cases. The extensive external validation of the influenza developed model in early COVID-19 cases will robustly demonstrate the performance in COVID-19 patients and show areas that need adjustment and the model's limitations. The lessons learned from this study could be used to inform the development of early prediction models in future pandemics.

METHODS

We performed a retrospective cohort study to develop COVID-19 prediction models for severe and critical illness. This study is reported according to the Transparent Reporting of a multivariate prediction model for Individual Prediction or Diagnosis (TRIPOD) guidelines(7).

At the start of the pandemic, there was very limited data available to develop prediction models due to the novel nature of the disease. To overcome the shortcoming of small data, we investigated whether we could use a proxy disease to develop a prediction model. This allowed us to utilise all available COVID-19 data for model validation. We developed models using historical data from patients with influenza or flu-like symptoms to assess a patient's individual risk of developing severe or critical illness following infection using readily available information (i.e. socio-demographics and medical history). The developed models were validated against COVID-19 patients to test whether the performance transferred between the two settings.

We developed COVID-19 Estimated Risk (COVER) scores to quantify a patient's risk of hospital admission with pneumonia (COVER-H), hospitalization with pneumonia requiring intensive services or death (COVER-I), or fatality (COVER-F) due to COVID-19 using the Observational Health Data Sciences and Informatics (OHDSI) Patient-Level Prediction framework(8). The research collaboration known as OHDSI has developed standards and tools that allow patient-level prediction models to be rapidly developed and externally validated following accepted best practices(9). This allows us to overcome two shortcomings of previous COVID-19 prediction papers by reporting according to open science standards and implementing widespread external validation.

Source of data

This study used observational healthcare databases from six different countries. All datasets used in this paper were mapped into the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM)(10). The OMOP-CDM was developed for researchers to have diverse datasets in a consistent structure and vocabulary. This enables analysis code and software to be shared among researchers, which facilitates replication and external validation of the prediction models.

The OMOP-CDM datasets used in this paper are listed in Table 1. All COVID-19 data was collected prior to 4/27/2020.

Participants

For model development, we identified patients aged 18 or older with a general practice (GP), emergency room (ER), or outpatient (OP) visit with influenza or flu-like symptoms (fever and either cough, shortness of breath, myalgia, malaise, or fatigue), at least 365 days of prior observation time, and no symptoms in the preceding 60 days. The initial healthcare provider interaction was used as index date, which is the point in time a patient enters the cohort.

Table 1 Data sources formatted to the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) used in this research (data type: claims, electronic health/medical records (EHR/EMR), general practitioner (GP))

Database	Database Acronym	Country	Data type	Contains COVID-19 data?	Time period
Columbia University Irving Medical Center Data Warehouse	CUIMC	United States	EMR	Yes	Influenza: 1990-2020 COVID-19: March-April 2020
Health Insurance and Review Assessment	HIRA	South Korea	Claims	Yes	COVID-19: 1st January-4th April 2020
The Information System for Research in Primary Care	SIDIAP	Spain	GP and hospital admission EHRs linked	Yes	Influenza: 2006-2017 COVID-19: March 2020
Tufts Research Data Warehouse	TRDW	United States	EMR	Yes	Influenza: 2006-2020 COVID-19: March 2020
Department of Veterans Affairs	VA	United States	EMR	Yes	Influenza: 2009-2010, 2014-2019 COVID-19: 1st March-20th April
Optum© De-Identified ClinFormatics® Data Mart Database*	ClinFormatics	United States	Claims	No	2000-2018
Ajou University School of Medicine Database	AUSOM	South Korea	EHR	No	1996 - 2018
Australian Electronic Practice based Research Network	AU-ePBRN	Australia	GP and hospital admission EHRs linked	No	2012-2019
IBM MarketScan® Commercial Database	CCAE	United States	Claims	No	2000-2018
Integrated Primary Care Information	IPCI	Netherlands	GP	Yes	2006-2020
Japan Medical Data Center	JMDC	Japan	Claims	No	2005-2018
IBM MarketScan® Multi-State Medicaid Database	MDCD	United States	Claims	No	2006-2017
IBM MarketScan® Medicare Supplemental Database	MDCR	United States	Claims	No	2000-2018
Optum© de-identified Electronic Health Record Dataset	Optum EHR	United States	EHR	No	2006-2018

*Development database

For validation in COVID-19 we used a cohort of patients presenting at an initial healthcare provider interaction with a GP, ER, or OP visit with COVID-19 disease. COVID-19 disease was identified by a diagnosis code for COVID-19 or a positive test for the SARS-COV-2 virus that was recorded after 1/1/2020. We required patients to be aged 18 or over, have at least 365 days of observation time prior to the index date and no diagnosis of influenza, flu-like symptoms, or pneumonia in the preceding 60 days.

Outcome

We investigated three outcomes: 1) hospitalization with pneumonia from index up to 30 days after index, 2) hospitalization with pneumonia that required intensive services (ventilation, intubation, tracheotomy, or extracorporeal membrane oxygenation) or death after hospitalization with pneumonia from index up to 30 days after index, and 3) death from index up to 30 days after index. Note that death is included in the second outcome to avoid incorrectly classifying patients who died without receiving intensive services as "low risk".

The analysis code used to construct the participant cohorts and outcomes used for development and validation can be found in the R packages located at: https://github.com/ohdsi-studies/Covid19PredictionStudies

Sensitivity analyses

We performed sensitivity analyses which involved using different versions of the COVID-19 cohort with varying sensitivities and specificities. At the beginning of the pandemic less testing capacity was available and as such we wanted to try broader definitions. Hence, we investigated three additional definitions where we included patients with symptoms, influenza, and visits any time prior to 2020. We then performed identical analysis with these changed cohorts.

Predictors

We developed a data-driven model using age in groups (18-19, 20-25, 26-30, …, 95+), sex, and binary variables indicating the presence or absence of recorded conditions and drugs any time prior to the index date. Missing records are thus effectively imputed as zero, exceptions are age and sex, which are always recorded in the OMOP-CDM. In total, we derived 31,917 candidate predictors indicating the presence of unique conditions/drugs recorded prior to the index date (GP, ER, or OP visit) for each patient. When using a data-driven approach to model development, generally the resulting models contain many predictors. This may optimise performance, but can be a barrier to clinical implementation. The utility of models for COVID-19 requires that they can be widely implemented across worldwide healthcare settings. Therefore, in addition to a data-driven model, we investigated two models that include fewer candidate predictors.

The age/sex model used only age groups and sex as candidate predictors. The COVER scores used a reduced set of variables, which were obtained by the following process:

1. Multiple clinicians inspected the data-driven model to identify variables that had a high standardized mean difference between patients with and without the outcome calculated using the following equation:

   $$(standardisedMeanDifference = \frac{mean\ with\ outcome - mean\ witho\quad outcome}{\sqrt{variance\ with\ outcome + vari\qquad without\ outcome}})$$

   There are often multiple predictors which are related and correlated selected by the model, for example a model might select a condition occurrence in different time periods predating the index date. This could be simplified to one predictor saying only "Patient had condition X in history", instead of having multiple predictors specifying in which time period the condition occurred. Likewise, multiple codes that are probably related to a specific condition could be simplified in one predictor. We identified general categories from these such as 'heart disease' and 'diabetes'.
2. Phenotype definitions for each category were created. This was performed to make the definitions clinically meaningful.
3. We trained a LASSO logistic regression model on the original data using age groups, sex and the newly created predictors indicating whether the patient had any of the category predictors.
4. The coefficients of this reduced variable model were then multiplied by 10 and rounded to the nearest integer. This was done to make the model simpler to calculate.
5. This gave us the simple score-based model.

Sample size

The models were developed using the Optum© De-Identified ClinFormatics® Data Mart Database. We identified 7,344,117 valid visits with influenza or flu-like symptoms, of which 4,431,867 were for patients aged 18 or older, 2,977,969 of these had at least 365 days of prior observation time, and 2,082,277 of these had no influenza/symptoms/pneumonia in the 60 days prior to index. We selected a random sample of 150,000 patients from the total population, as research showed it is possible to efficiently develop models with near optimal performance, while reducing model complexity and computational requirements by using a sample of this size(45). Riley et al. provide a calculator for minimum sample size, which for number of predictors = 20, event rate = 0.05 and R^2 = 0.1 would require a minimum of 1,698 patients(46). This subset was used to develop the data-driven model. The full set of 2,082,077 patients was then used for the development and validation of the simple model. A small subset of this data was used to develop the data-driven model and so the presented internal performance could be optimistic. In theory this is a limitation, but it has no effect on the evidence of the external validation. Figure 1 is a flow chart demonstrating the above exclusions and flow of data through the study.

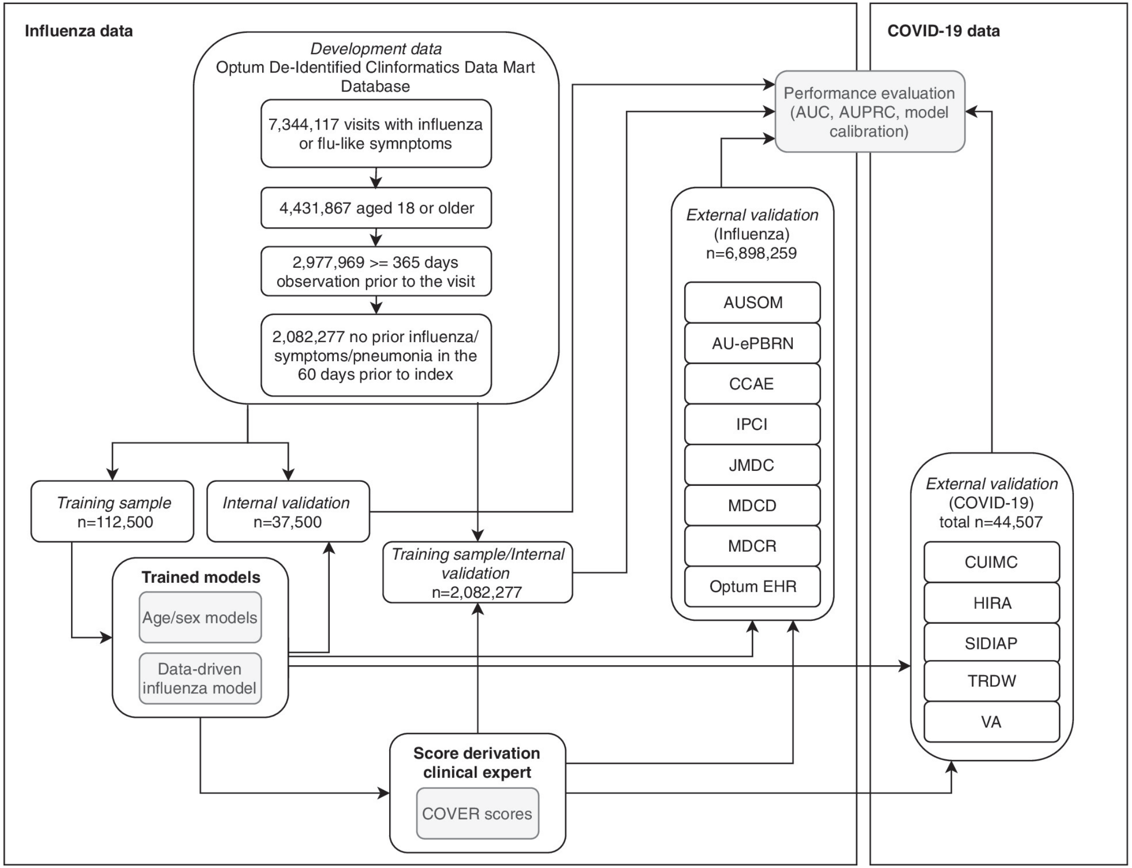

Influenza data
COVID-19 data
Development data
Optum De-Identified Clinformatics Data Mart Database
7,344,117 visits with influenza or flu-like symnptoms
4,431,867 aged 18 or older
2,977,969 >= 365 days observation prior to the visit
2,082,277 no prior influenza/ symptoms/pneumonia in the 60 days prior to index
Training sample n=112,500
Internal validation n=37,500
Trained models
Age/sex models
Data-driven influenza model
Training sample/Internal validation n=2,082,277
Score derivation clinical expert
COVER scores
External validation (Influenza) n=6,898,259
AUSOM
AU-ePBRN
CCAE
IPCI
JMDC
MDCD
MDCR
Optum EHR
Performance evaluation (AUC, AUPRC, model calibration)
External validation (COVID-19) total n=44,507
CUIMC
HIRA
SIDIAP
TRDW
VA

Figure 1 A Flow chart representing the path of data in the study. This details the splits used internally for model development, the steps taken for model parsimonisation and validation and external validation

Missing data

Age and sex are required by the OMOP-CDM used by OHDSI and will never be missing. For each condition or drug we considered no records in the database to mean the patient does not have the condition or does not receive the drug. This could lead to misclassification of patients if a patient's illness is not recorded in the database.

Statistical analysis methods

Model development followed a previously validated and published framework for the creation and validation of patient-level prediction models(8). We used a person 'train-test split' method to perform internal validation. In the development cohort, a random split sample (`training sample') containing 75% of patients was used to develop the prediction models and the remaining 25% of patients (`test sample') was used to internally validate the models. We trained models using LASSO regularized logistic regression, using a 3-fold cross validation technique in the train-set to learn the optimal regularization hyperparameter through an adaptive search(13). We used R (version 3.6.3) and the OHDSI Patient-Level Prediction package (version 3.0.16) for all statistical analyses(8).

To evaluate the performance of the developed models, we calculate the overall discrimination of the model using the area under the receiver operating characteristic curve (AUC), the area under the precision recall-curve (AUPRC), and the model calibration. The AUC indicates the probability that for two randomly selected patients, the patient who gets the outcome will be assigned a higher risk. The AUPRC shows the trade-off between identifying all patients who get the outcome (recall) versus incorrectly identifying patients without outcome (precision) across different risk thresholds. The model calibration is presented in a plot to examine agreement between predicted and observed risks across deciles of predicted risk. Calibration assessment is then performed visually rather than using a statistic or numeric value as this provides a better impression of the direction and scale of miscalibration(14). Summary statistics are reported from the test samples.

We performed external validation in databases containing COVID-19 data. To do this we assessed patients with confirmed COVID-19. In addition, we performed a classical external validation in which we applied the models to identical settings across diverse patient populations with influenza or flu-like symptoms prior to 2020. We examined the external validation using AUC, AUPRC and model calibration in the same way as internally. We provide confidence intervals when the number of events is below 1,000. Once the number of events increases, confidence intervals become too narrow to provide a good estimate of error.

This study adheres to open science principles for publicly prespecifying and tracking changes to study objectives, protocol, and code as described in the Book of OHDSI(15). For transparency, the R packages for the development and external validation of the models in any database mapped to the OMOP-CDM are available on GitHub at: https://github.com/ohdsi-studies/Covid19PredictionStudies

RESULTS

Online results

The complete results are available as an interactive app at: http://evidence.ohdsi.org/Covid19CoverPrediction

This application will continue to be updated as the models are validated, an archived version of the app that was released to accompany this article is available here: https://zenodo.org/record/4697417

Participants

Table 2 describes the characteristics at baseline of the patients across the databases used for development and external validation. Out of the 150,000 patients sampled with influenza or flu-like symptoms in the development database (ClinFormatics), there were 6,712 patients requiring hospitalization with pneumonia, 1,828 patients requiring hospitalization and intensive services with pneumonia or death, and 748 patients died within 30 days. See Table 2 for the full outcome proportions across the databases included in this study. A total of 44,507 participants with COVID-19 disease were included for external validation.

In the databases used for external validation, the patient numbers ranged from 395 (TRDW) to 3,146,743 (CCAE). The datasets had varied outcome proportions ranging from 0.06-12.47 for hospital admission, 0.01-4.91 for intensive services, and 0.01-12.27 for fatality. Characteristics at baseline differed substantially between databases as can be seen in Table 2, with MDCR (a database representing retirees) containing a relatively old population of patients and a high number of comorbidities, and IPCI (a database representing general practice) showing a relatively low condition occurrence.

Model performance

The internal validation performance for each model is presented in Table 3. The external validation of the COVER scores on the COVID-19 patients is shown in Table 4. Full validation results can be seen in Appendix 1B of the online supplement. Receiver operating characteristic and calibration plots are included in Figure 2.

Model specification

The data-driven models for hospitalization, intensive services, and fatality contained 521, 349, and 205 predictors respectively. The COVER-H, COVER-I, and COVER-F scores are presented in Figure 3. After data-driven selection, clinicians reviewed the resulting models and created the composite predictors. This produced the COVER scores which include 7 predictors, in addition to age groups and sex, that corresponded to the following conditions existing any time prior to the index date: cancer, chronic obstructive pulmonary disease, diabetes, heart disease, hypertension, hyperlipidemia, and kidney disease (chronic and acute). A description of

Table 2 Population size, outcome proportion, and characteristics for the development database (influenza) and external validation databases for COVID-19 and influenza (N/A indicates this result is not available)

	Development	**External validation: COVID-19**					**External validation: influenza**							
	ClinFormatics	CUIMC	HIRA	SIDIAP	TRDW	VA	AUSOM	AU-ePBRN	CCAE	IPCI	JMDC	MDCD	MDCR	Optum EHR
Number of participants	2,082,277	2,731	1,985	37,950	395	1,446	3,105	2,791	3,146,801	29,132	1,276,478	536,806	248,989	1,654,157
Hospitalization with pneumonia (Outcome proportion %)	105,030 (5.04)	N/A	89 (4.48)	1,223 (1.11)	21 (5.32)	149 (10.30)	49 (1.58)	29 (1.04)	33,824 (1.07)	22 (0.08)	728 (0.06)	32,987 (6.15)	31,059 (12.47)	34,229 (2.07)
Hospitalization with pneumonia requiring intensive services or death (Outcome proportion %)	29,905 (1.44)	134 (4.91)	22 (1.11)	N/A	5 (1.27)	38 (2.63)	5 (0.16)	3 (0.11)	4,856 (0.02)	24 (0.08)	65 (0.01)	7,226 (1.35)	3,628 (1.46)	7,368 (0.45)
Death (Outcome proportion %)	11,407 (0.55)	335 (12.27)	43 (2.17)	406 (1.07)	1 (0.25)	43 (2.97)	5 (0.16)	4 (0.14)	965 (0.03)	24 (0.08)	75 (0.01)	2,603 (0.48)	1,354 (0.54)	3,513 (0.21)
Age (% above 65)	26.1	38.9	15.6	17.9	18.2	37.3	11.9	23.1	12.5	16.9	16.0	14.2	96.2	30.0
Sex (%, male)	44.4	47.2	43.5	43.4	49.6	81.4	41.7	44.5	42.7	43.7	56.8	29.2	45.9	40.1
Cancer (%)	12.6	17.1	9.8	6.3	11.6	17.0	7.7	8.2	6.2	3.7	2.5	8.9	35.2	10.6
COPD (%)	10.2	9.3	4.9	2.5	6.3	20.5	2.7	3.1	2.7	2.7	0.5	19.8	26.6	7.6
Diabetes (%)	20.5	30.9	23.1	8.0	19.7	35.2	3.8	13.0	11.4	6.7	8.3	27.4	36.1	15.3
Heart disease (%)	31.0	40.1	17.1	11.2	25.8	44.7	7.7	12.9	16.5	7.5	8.0	36.1	68.2	23.4
Hypertension (%)	44.2	51.6	26.3	14.8	38.5	63.0	13.9	27.0	29.1	12.4	11.4	49.8	80.4	36.1
Hyperlipidemia (%)	46.8	40.6	39.9	11.4	32.9	62.5	3.3	20.2	21.8	4.6	15.2	36.0	69.6	34.2
Kidney disease (%)	18.7	31.2	17.0	11.0	24.3	32.4	7.6	6.2	9.0	1.2	5.1	23.4	35.5	14.9

Table 3 Results for internal validation in ClinFormatics

Outcome	Predictors	No. Variables	AUC	AUPRC
Hospitalization with pneumonia	Conditions/drugs + age/sex	521	0.852	0.224
	Age/sex	2	0.818	0.164
	COVER-H	9	0.840	0.120
Hospitalization with pneumonia requiring intensive services or death	Conditions/drugs + age/sex	349	0.860	0.070
	Age/sex	2	0.821	0.049
	COVER-I	9	0.839	0.059
Fatality	Conditions/drugs + age/sex	205	0.926	0.069
	Age/sex	2	0.909	0.037
	COVER-F	9	0.896	0.039

Table 4 Results of external validation of the COVER scores on COVID-19 patients with a GP, ER, or OP visit in 2020 (*Confidence interval is not reported as the number of outcomes is larger than 1000)

Outcome	Database	AUC (95% confidence interval)	AUPRC
Hospitalization with pneumonia (COVER-H)	HIRA	0.806 (0.762-0.851)	0.134
	SIDIAP	0.748*	0.072
	TRDW	0.731 (0.611-0.851)	0.132
	VA	0.689 (0.649-0.729)	0.179
Hospitalization with pneumonia requiring intensive services or death (COVER-I)	CUIMC	0.734 (0.699-0.769)	0.100
	HIRA	0.910 (0.889-0.931)	0.053
	VA	0.763 (0.708-0.818)	0.058
Fatality (COVER-F)	CUIMC	0.820 (0.796-0.840)	0.400
	HIRA	0.898 (0.857-0.940)	0.150
	SIDIAP	0.895 (0.881-0.910)	0.083
	VA	0.717 (0.642-0.791)	0.068

the covariates can be found in Appendix 1A of the online supplement. The COVER scores are detailed in Figure 3 and are accessible online under the calculator tab at: http://evidence.ohdsi.org:3838/Covid19CoverPrediction/

Figure 3 also provides a risk converter, which allows for easy conversion between the risk score and predicted risk of the outcomes. The scores can be converted to a probability by applying the logistic function: 1/(1+exp((risk score-93)/10)). Furthermore, we provide a plot of the probability distribution for each of the three models from patients in ClinFormatics to demonstrate the expected regions the probabilities fall into. To calculate the COVER scores using Figure 3, a clinician first needs to identify which conditions the patient has. The points for the corresponding predictors are then added to arrive at the total score. For example, if a 63-year-old female patient has diabetes and heart disease, then her risk score for hospital admission (COVER-H) is 43 (female sex) + 4 (heart disease) + 3 (diabetes) + 15 (age) = 65. The risk scores for intensive services (COVER-I) and fatality (COVER-F) are 51 and 47, respectively.

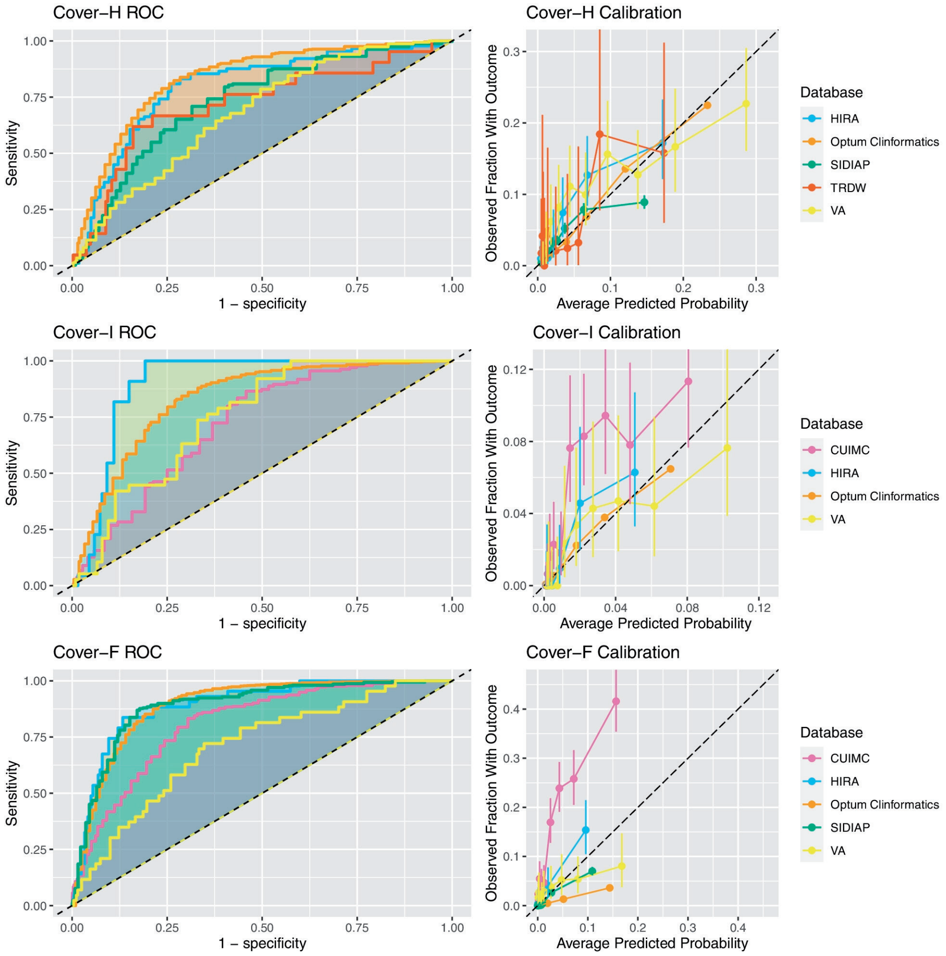


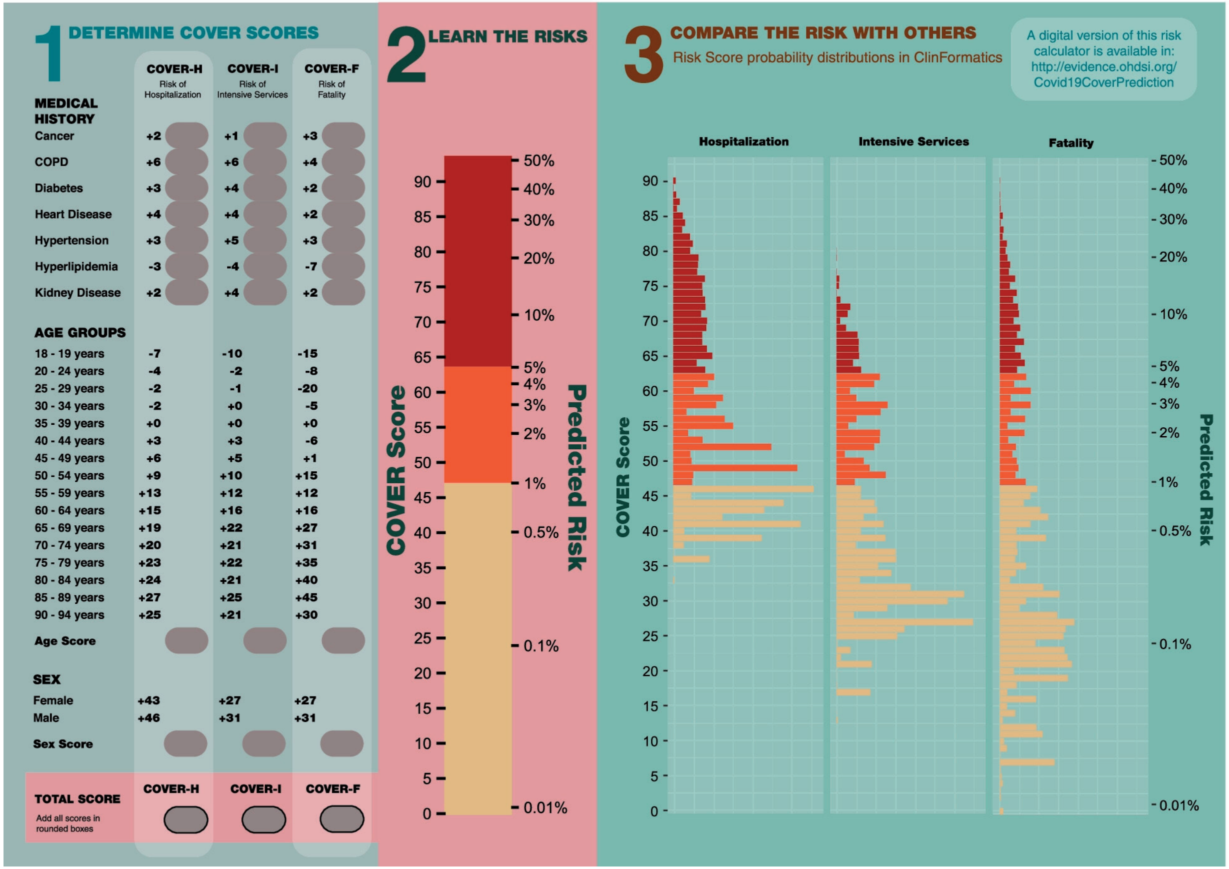
Figure 2 The ROC and Calibration plots for the validations (internal and external) of the 3 Cover scores

Using the risk converter in Figure 3, a score of 65 corresponds to a risk of 6%. Scores of 51 and 47 correspond to 1.5% and 1%, respectively.

DISCUSSION

Interpretation

We developed and externally validated models using large datasets of influenza patients to quantify a patient's risk of developing severe or critical illness due to COVID-19. In the development data, the 9-predictor COVID-19 Estimated Risk (COVER) scores were a good trade-off between model complexity and performance, as the AUCs were generally close to the large data-driven models. In the development database the COVER scores achieved an AUC of



1 DETERMINE COVER SCORES

	COVER-H Risk of Hospitalization	COVER-I Risk of Intensive Services	COVER-F Risk of Fatality
MEDICAL HISTORY			
Cancer	+2	+1	+3
COPD	+6	+6	+4
Diabetes	+3	+4	+2
Heart Disease	+4	+4	+2
Hypertension	+3	+5	+3
Hyperlipidemia	-3	-4	-7
Kidney Disease	+2	+4	+2
AGE GROUPS			
18 - 19 years	-7	-10	-15
20 - 24 years	-4	-2	-8
25 - 29 years	-2	-1	-20
30 - 34 years	-2	+0	-5
35 - 39 years	+0	+0	+0
40 - 44 years	+3	+3	-6
45 - 49 years	+6	+5	+1
50 - 54 years	+9	+10	+15
55 - 59 years	+13	+12	+12
60 - 64 years	+15	+16	+16
65 - 69 years	+19	+22	+27
70 - 74 years	+20	+21	+31
75 - 79 years	+23	+22	+35
80 - 84 years	+24	+21	+40
85 - 89 years	+27	+25	+45
90 - 94 years	+25	+21	+30
Age Score			
SEX			
Female	+43	+27	+27
Male	+46	+31	+31
Sex Score			
TOTAL SCORE Add all scores in rounded boxes	COVER-H	COVER-I	COVER-F

2 LEARN THE RISKS
COVER Score
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
0
Predicted Risk
50%
40%
30%
20%
10%
5%
4%
3%
2%
1%
0.5%
0.1%
0.01%
3 COMPARE THE RISK WITH OTHERS
Risk Score probability distributions in ClinFormatics
A digital version of this risk calculator is available in: http://evidence.ohdsi.org/Covid19CoverPrediction
Hospitalization
Intensive Services
Fatality
COVER Score
Predicted Risk

Figure 3 A graphic showing how to calculate the 3 Cover scores with a nomogram to convert the raw score into a percentage risk. There is also a distribution of scores found using internal validation to allow for comparison of a patients score to the wider populat

0.84 when predicting which patients will be hospitalized or require intensive services and an AUC of 0.90 when predicting which patients will die within 30 days. When validated on 1,985 COVID-19 patients in South Korea the COVER-H score achieved an AUC of 0.81, COVER-I and COVER-F achieved an AUC of 0.90 and 0.91. When applied to 37,950 COVID-19 Spanish patients COVER-H had an AUC of 0.75 and performed better when predicting fatality (COVER-F: AUC 0.89). When applied to US patients, the COVER-I and COVER-F models achieved AUCs of 0.73 and 0.82 in CUIMC, VA performed similarly with AUCs of 0.76 and 0.72 respectively. The VA also achieved 0.69 for COVER-H. The results show reasonable performance with some inconsistency across a range of countries.

A visual assessment of calibration plots across validations showed reasonable calibration in HIRA, SIDIAP, and VA. There was a slight overestimation of risk amongst oldest and highest risk strata in SIDIAP, and to a lesser extent in HIRA. The calibration was poor in CUIMC, as risk was often underestimated. This may be due to CUIMC containing mostly hospitalized COVID-19 patients, so the CUIMC cohort are experiencing more severe COVID-19. The VA showed some miscalibration in the lowest and highest risk strata. The observed miscalibration is possibly due to the differing severities of the diseases used for model development and calibration. However, miscalibration could also be due to other differences in populations not caused by the use of a proxy disease. The variable calibration results suggest that the model's performance should be assessed and the model should potentially be recalibrated before being implemented in a local context. A simple method to do this is by adjusting the baseline risk based upon the differences found between development and validation populations using an adjustment factor derived from the differences in case mix between development and validation settings(16, 17).

The age/sex models also show reasonable performance, and these predictors are among the main contributors to performance in the COVER scores. This suggests these models could also be suitable if access to medical history is difficult.

These results showed that training in large historical influenza data was an effective strategy to develop models for COVID-19 patients. We also performed sensitivity analyses using more sensitive COVID-19 definitions, for example including patients with symptoms, influenza, and visits any time prior to 2020. The results did not show much deviation from the specific definition (online supplement Appendix 1B). Our results show that quantifying a symptomatic patient's risk based on a small selection of comorbidities as well as age/sex gives improved model performance.

Limitations

First, it has become clear that there are differences in the underlying nature of the two diseases, particularly in respect to the severity of symptoms in COVID-19 patients compared with influenza patients. Therefore, it is possible another disease may have provided a better proxy than influenza.

Second, despite preserving all the target disease data for validation, we still had relatively low outcome numbers. In the CUIMC, HIRA, SIDIAP, and VA COVID-19 databases we either reached or approached the threshold for reliable external validation of 100 patients who experience the outcome of interest(18, 19), but the results of TRDW might not be reliable.

Furthermore, the data reported early during the COVID-19 pandemic was noisy and skewed. This might cause misclassification in the target and outcome cohorts. In order to counter this, we performed sensitivity analysis using cohorts that included broad and narrow COVID-19 definitions, the impact of this on the results was minimal. The use of a 30-day risk window has the limitation that if a patient experiences an outcome after the time window, this will be (incorrectly) recorded as a non-event. There is further potential misclassification of predictors, for example, if a disease is incorrectly recorded in a patient's history. Moreover, the result of the phenotype generation process is not fully reproducible due to the use of clinician expertise, which is an unresolved problem in much epidemiological work. However, the phenotype development process is reproducible and the phenotypes generated are provided. The evidence in the paper shows the models to be robust and transportable.

We were unable to include some suspected disease predictors in the analysis as these are not readily available (e.g. lymphocyte count, lung imaging features) or inconsistently collected and reported across the various databases included in the study (e.g. BMI, ethnicity). However, due to the high load on healthcare systems and the contagious nature of the disease we believe it is useful to have a model that does not require a patient to be either in hospital or another setting to receive tests. A similar issue also meant we were not able to validate the COVER-H score in CUIMC (it mostly contains ER or hospitalized COVID-19 patients) and the COVER-I score in SIDIAP (due to a lack of information on intensive services in the database).

Finally, concerns exist over the clinical validity of claims data, however we were able to develop models using claims data that transported well into EHR data. There is the potential for some overlap of patients between claims and EHR databases, although this number is likely to be small.

Implications

The results show we were able to develop models that use historical influenza patient's socio-demographics and medical history to predict their risk of becoming severely or critically ill when infected with COVID-19. To our knowledge, this is the first study that has been able to extensively externally validate prediction models on COVID-19 patients at a global scale. The adequate performance of the COVER scores in COVID-19 patients (as quantified by consistent finding of AUC > 0.7 in new settings) show these scores could have been used to identify patients who should have been shielded from COVID-19 in the early stages of the pandemic.

CONCLUSION

In this paper we developed and validated models that can predict which patients presenting with COVID-19 are at high risk of experiencing severe or critical illness. This research demonstrates that it is possible to develop a prediction model rapidly using historical data of a similar disease that, once re-calibrated with contemporary data and outcomes from the current outbreak, could be used to help inform strategic planning and healthcare decisions.

Supplementary Information

The online version contains supplementary material available at https://rdcu.be/c7aDU

Declarations

Ethics approval and consent to participate

The manuscript uses secondary data and as such no human participants were involved in the study and informed consent was not necessary at any site.

Consent for publication

Not applicable

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available due to patient privacy and data protection concerns. Information on access to the databases is available from the corresponding author.

Competing interests

RDW, AM, CY, TDS, TF, JJ, CK, YR, AW, AA, MHO, MA, CA, EB, YHC, ID, MTFA, SFB, JAK, STL, FN, RWP, NP, TS, MS, EWS, SCY, LZ(Zhang), LZ(Zhou), report no conflicts of interest.

MM reports grants from Department of Veteran Affairs, grants from National Institutes of Health, outside the submitted work. SLD reports grants from Anolinx, LLC, grants from Astellas Pharma, Inc, grants from AstraZeneca Pharmaceuticals LP, grants from Boehringer Ingelheim International GmbH, grants from Celgene Corporation, grants from Eli Lilly and Company, grants from Genentech Inc., grants from Genomic Health, Inc., grants from Gilead Sciences Inc., grants from GlaxoSmithKline PLC, grants from Innocrin Pharmaceuticals Inc., grants from Janssen Pharmaceuticals, Inc., grants from Kantar Health, grants from Myriad Genetic Laboratories, Inc., grants from Novartis International AG, grants from Parexel International Corporation, outside the submitted work. BK-H reports grants from Innovation Fund Denmark, grants from Novo Nordisk Foundation, during the conduct of the study. DPA reports grants and other from

UCB Biopharma SRL, grants and other from Amgen, grants from Servier Laboratoires, outside the submitted work. DM reports grants from Wellcome Trust (Clinical Research Development Fellowship), grants from Chief Scientist Office (CSO), grants from Health Data Research UK (HDR-UK), grants from National Institute of Health Research (NIHR), outside the submitted work. AP-U reports grants from Fundacion Alfonso Martin Escudero, grants from Medical Research Council, outside the submitted work. GH reports grants from NIH National Library of Medicine, during the conduct of the study. grants from Janssen Research, outside the submitted work. JMR, PBR, AS, GR are employees of Janssen Research and Development and are shareholders of Johnson & Johnson. KK and CR report as employees of IQVIA. MR reports personal fees from Bayer Hispania, S.L., outside the submitted work. PLK and LZ are AbbVie employees. MAS reports grants from US National Institutes of Health, grants from IQVIA, personal fees from Janssen Research and Development, personal fees from Private Health Management, during the conduct of the study. PRR reports grants from Innovative Medicines Initiative, grants from Janssen Research and Development, during the conduct of the study. GM is a full-time employee of Janssen Latin America, outside the submitted work;.

Funding

This project has received support from the European Health Data and Evidence Network (EHDEN) project. EHDEN received funding from 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. This project is funded by the Health Department from the Generalitat de Catalunya with a grant for research projects on SARS-CoV-2 and COVID-19 disease organized by the Direcció General de Recerca i Innovació en Salut. The University of Oxford received a grant related to this work from the Bill & Melinda Gates Foundation (Investment ID INV-016201), and partial support from the UK National Institute for Health Research (NIHR) Oxford Biomedical Research Centre. DPA is funded through a NIHR Senior Research Fellowship (Grant number SRF-2018-11-ST2-004). The views expressed in this publication are those of the author(s) and not necessarily those of the NHS, the National Institute for Health Research, or the Department of Health. AP-U is supported by Fundacion Alfonso Martin Escudero and the Medical Research Council (grant numbers MR/K501256/1, MR/N013468/1). BSKH is funded through Innovation Fund Denmark (5153-00002B) and the Novo Nordisk Foundation (NNF14CC0001). This work was also supported by the Bio Industrial Strategic Technology Development Program (20001234) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea) and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number: HI16C0992]. This project is part funded by the UNSW RIS grant. This research received funding support from the US Department of Veterans Affairs and the VA Informatics and Computing Infrastructure (VA

HSR RES 13-457). The views and opinions expressed are those of the authors and do not necessarily reflect those of the Department of Veterans Affairs or the United States Government.

Authors' contributions

All authors made substantial contributions to the conception or design of the work; JMR and RDW led the data analysis; all authors were involved in the analysis and interpretation of data for the work; all authors have contributed to the drafting and revising critically the manuscript for important intellectual content; all authors have given final approval and agree to be accountable for all aspects of the work.

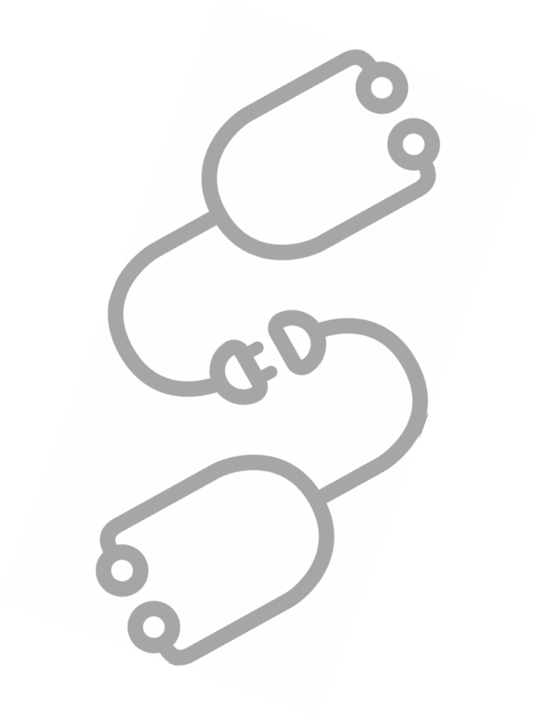
Acknowledgements

The authors would like to thank the OHDSI community for their contributions to the tools used for this analysis.

The authors appreciate healthcare professionals dedicated to treating COVID-19 patients in Korea, and the Ministry of Health and Welfare and the Health Insurance Review & Assessment Service of Korea for sharing invaluable national health insurance claims data in a prompt manner.

BIBLIOGRAPHY

1. World Health Organization. *Clinical management of severe acute respiratory infection* (SARI) when COVID-19 disease is suspected: interim guidance, 13 March 2020. Geneva:World Health Organization;2020.
2. Prieto-Alhambra D, Ballo E, Coma-Redon E, et al. Hospitalization and 30-day fatality in 121,263 COVID-19 outpatient cases. *medRxiv.* 2020:2020.2005.2004.20090050.
3. World Health Organization. *Coronavirus disease 2019 (COVID-19) Situation report - 51 2020, 11 March 2020.* World Health Organization;2020.
4. Wynants L, Van Calster B, Bonten MMJ, et al. Prediction models for diagnosis and prognosis of covid-19 infection: systematic review and critical appraisal. *BMJ.* 2020;369:m1328.
5. Piroth L, Cottenet J, Mariet AS, et al. Comparison of the characteristics, morbidity, and mortality of COVID-19 and seasonal influenza: a nationwide, population-based retrospective cohort study. *Lancet Respir Med.* 2021;9(3):251-259.
6. Petersen E. COVID-19 is not influenza. *Lancet Respir Med.* 2021;9(3):219-220.
7. Moons KG, Altman DG, Reitsma JB, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. *Ann Intern Med.* 2015;162(1):W1-73.
8. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *J Am Med Inform Assoc.* 2018;25(8):969-975.
9. Reps JM, Williams RD, You SC, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. *BMC Med Res Methodol.* 2020;20(1):102-102.
10. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc.* 2012;19(1):54-60.
11. John LH, Kors JA, Reps JM, Ryan PB, Rijnbeek PR. How little data do we need for patient-level prediction? *arXiv preprint arXiv:200807361.* 2020.
12. Riley RD, Ensor J, Snell KIE, et al. Calculating the sample size required for developing a clinical prediction model. *BMJ.* 2020;368:m441.
13. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D. Massive parallelization of serial inference algorithms for a complex generalized linear model. *ACM Transactions on Modeling and Computer Simulation (TOMACS).* 2013;23(1):1-17.
14. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J.* 2014;35(29):1925-1931.
15. Observational Health Data Sciences and Informatics. *The Book of OHDSI.* 2019.
16. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med.* 2004;23(16):2567-2586.
17. Janssen KJ, Moons KG, Kalkman CJ, Grobbee DE, Vergouwe Y. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol.* 2008;61(1):76-86.
18. Vergouwe Y, Steyerberg EW, Eijkemans MJ, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. *J Clin Epidemiol.* 2005;58(5):475-483.
19. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. *Stat Med.* 2016;35(2):214-226.



3

Feasibility and Evaluation of a Large-Scale External Validation Approach for Patient-Level Prediction in an International Data Network: Validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation

Jenna M. Reps[1], Ross D. Williams[2], Seng Chan You[3], Thomas Falconer[4], Evan Minty[5], Alison Callahan[6] Patrick B. Ryan[1], Rae Woong Park[3,7], Hong-Seok Lim[8], Peter Rijnbeek[2]

[1]Janssen Research and Development, Titusville, NJ; [2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands, [3]Department of Biomedical informatics, Ajou University School of Medicine, Suwon, Republic of Korea; [4]Department of Biomedical Informatics, Columbia University Medical Center, New York; [5]O'Brien Institute for Public Health, Faculty of Medicine, University of Calgary, Calgary, Alberta, Canada; [6]Center for Biomedical Informatics Research, School of Medicine, Stanford University, Stanford CA; [7]Department of Biomedical Sciences, Ajou University Graduate School of Medicine, Suwon, Republic of Korea; [8]Department of Cardiology, Ajou University Medical Centre, Suwon, Republic of Korea

Link to article: https://rdcu.be/digCJ

ABSTRACT

Background: To demonstrate how the Observational Healthcare Data Science and Informatics (OHDSI) collaborative network and standardization can be utilized to scale-up external validation of patient-level prediction models by enabling validation across a large number of heterogeneous observational healthcare datasets.
Methods: Five previously published prognostic models (ATRIA, CHADS$_2$, CHADS$_2$VASC, Q-Stroke and Framingham) that predict future risk of stroke in patients with atrial fibrillation were replicated using the OHDSI frameworks. A network study was run that enabled the five models to be externally validated across nine observational healthcare datasets spanning three countries and five independent sites.
Results: The five existing models were able to be integrated into the OHDSI framework for patient-level prediction and they obtained mean c-statistics ranging between 0.57-0.63 across the 6 databases with sufficient data to predict stroke within 1 year of initial atrial fibrillation diagnosis for females with atrial fibrillation. This was comparable with existing validation studies. The validation network study was run across nine datasets within 60 days once the models were replicated. An R package for the study was published at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ExistingStrokeRiskExternalValidation .
Conclusion: This study demonstrates the ability to scale up external validation of patient-level prediction models using a collaboration of researchers and a data standardization that enable models to be readily shared across data sites. External validation is necessary to understand the transportability or reproducibility of a prediction model, but without collaborative approaches it can take three or more years for a model to be validated by one independent researcher. In this paper we show it is possible to both scale-up and speed-up external validation by showing how validation can be done across multiple databases in less than 2 months. We recommend that researchers developing new prediction models use the OHDSI network to externally validate their models.

BACKGROUND

Observational healthcare data often contains longitudinal medical records for large heterogeneous populations. There has been increased interest in learning patient-level prediction models using these big real-world datasets with the aim of improving healthcare [1]. These patient-level prediction models can be used to identify high-risk subgroups that could benefit from interventions. For example, the cardiovascular model QRISK2, that was developed using a UK primary care database, is used to identify patients who may benefit from lipid -lowering medication [2]. It is important to ensure a model has good performance before it is used clinically and this requires external validation [1,3].

Models are often internally validated using the development dataset by withholding a subset of that data from the model training stage so that it can be used for evaluating the model performance. The majority of patient-level prediction models will report internal validation. External validation is accomplished by evaluating the model on a new dataset (that is different from the development dataset). Few published patient-level prediction models are externally validated, and research has shown that it often takes three or more years for external validation to occur once a model is published [4].

External validation of a patient-level prediction model can provide useful insights into the accuracy of the model across different patient characteristics and may be used to learn the impact of missing predictors. The type of external validation depends on the similarity between the development and validation datasets. When a model is validated on a population that has similar characteristics to the development data population the 'generalizability performance' of the model is investigated (i.e., how well the model performs when making predictions on similar patients). When a model is validated on a population that has different characteristics to the development data population the 'transportability performance' of the model is investigated (i.e., how well the model performs on different patients). Many observational datasets are not representative of the whole population, so the transportability performance of the model discovered during external validation on patients with different characteristics is important to know when identifying who the model can be broadly applied to. For example, some clinical guidelines recommend treatment stratification for patients based on applying a simple risk score model that was developed on a small population but the transportability of the model to the general population may not have been studied. This may lead to incorrect predictions.

External validation is a slow process due to the difficulty finding suitable data to replicate a prediction model on and difficulty replicating a prediction model (e.g., writing code to correctly extract the same model covariates from the new data). Often published papers lack the information required to replicate the model or can be interpreted subjectively (e.g., in defining medical conditions or variables) which can be an issue causing models to be replicated incorrectly. This prevents efficient and large-scale external validation which likely slows down clinical uptake of published patient-level prediction models or results in the models being applied clinically to patient populations where the model transportability is unknown.

A collaborative approach to external model validation has been proposed to enable extensive evaluation of prediction models [5]. The Observational Healthcare Data Science and Informatics (OHDSI) network is a community of researchers that are working towards the common goal of improving the analysis of observational data. The OHDSI community have developed standardizations that enable efficient collaboration across research sites. The main standardization is the common data structure and vocabulary used by all collaborators known as the Observational Medical Outcomes Partnership (OMOP) common data model. The OMOP common data model ensures all researchers have their data in the same structure so analysis codes such as Structured Query Language (SQL) can be shared across sites. This has enabled the development of analysis packages in R for causal inference and patient-level prediction that can be used by any researcher with data in the OMOP common data model. The OHDSI collaborative network, common data model and patient-level prediction package now present the opportunity to scale up external validation.

The aim of this study is to demonstrate that the OHDSI tools and OMOP common data model can be used by researchers to investigate the external validation performance of their prediction models across a large number of heterogeneous patient populations. Instead of taking years to externally validate a model, OHDSI may make it possible to apply a prediction models to a large number of datasets in a short period of time. To demonstrate this we selected the prediction problem of 1-year risk of stroke in newly diagnosed atrial fibrillation patients as there are multiple existing models that are used clinically, namely Anticoagulation and Risk Factors in Atrial Fibrillation (ATRIA) (no prior stroke model) [6], Framingham (no prior stroke model) [7] , Congestive heart failure, Hypertension, Age > 75, Diabetes, prior Stroke/transient ischemic attack (CHADS$_2$) [8], CHADS$_2$-VASc [9] and Q-Stroke (female model) [10]. We show these models can be replicated using the OHDSI standardizations and externally validated across numerous data sites within the OHDSI network.

METHODS

Existing stroke prediction models

We selected the problem of predicting stroke in patients with atrial fibrillation as it has been well studied and is one of the only prediction problems to have been extensively validated. Therefore, we have ample benchmarks to compare to the results of this study. The existing models we replicated were ATRIA, CHADS$_2$, CHA$_2$DS$_{2-}$VASc, Framingham and Q-Stroke.

The ATRIA [6] model was developed on a cohort of 7,284 patients who were 18+ and had an atrial fibrillation outpatient diagnosis during 1997 or 1998. ATRIA was internally validated on a 3,643 patient hold out set obtaining a c-statistic of 0.72. In the same paper, the authors also externally validated the model on a cohort of 33,247 patients aged 21+ with inpatient or outpatient atrial fib or flutter during 2006-2009, obtaining a c-statistic of 0.7. The CHADS$_2$ score [8] was developed by combining two other stroke prediction models (using the variables from these models and assigning points) and was validated on 1,733 patients aged 65 to 95 years who had nonrheumatic atrial fibrillation. The CHADS$_2$ score obtained a c-statistic of 0.81 on this population. The CHA$_2$DS$_{2-}$VASc score [9] is another score-based model that was developed using knowledge of risk factors. The model was validated on a cohort of 1,577 patients who were 18+ and had atrial fibrillation during 2003 to 2004 from 35 countries. The model obtained a c-statistic of 0.61 for this patient population. The Framingham score [7] model was based on a Cox model developed using data from 705 patients aged 55 to 94 with initial atrial fibrillation. The internal validation, using a bootstrap approach, showed a c-statistic of 0.66. The Q-Stroke [10] model was developed using primary care data from the UK consisting of 3,549,478 patients aged 25-84 with no prior stroke or anticoagulation use (except aspirin) and was internally validated on 1,897,168 similar patients. When applying the model to predict the 10-year risk of stroke in female patients with atrial fibrillation at baseline, the c-statistic was 0.65.

The existing models include a small number of variables, Table 1 summarizes the variables included in each model. Some of the variables are unlikely to be available in claims data and these are marked with the + symbol. A large number of Q-Stroke variables are not commonly recorded in claims data (or are UK specific), so this model is difficult to replicate in external non-UK databases. For example, US claims data contain incomplete measurement records and rarely record family history but many of the Q-stroke predictors were recent measurements or family history. Table 2 presents the internal performance and published external validation performance for the five models. Although the internal validation c-statistic for some of the models was as high as 0.8, independent external validation studies of the models tend to show the models achieve c-statistics between 0.6 and 0.7.

The complete definitions for each variable (sets of SNOMED CT or RXNorm codes) are provided in online Appendix A.

Validation Prediction task

Within a target population of female patients with newly diagnosed atrial fibrillation and no prior stroke predict who will develop a stroke 1 to 365 days after initial diagnosis of atrial fibrillation.

Sources of Data

We validated the existing models using a retrospective cohort design and various observational healthcare datasets (e.g., claims data and electronic healthcare data). The datasets used to evaluate the models are:

IBM MarketScan® Commercial Database (CCAE) is a United States employer-sponsored insurance health plans claims database. The database contains claims (e.g. inpatient, outpatient,

Table 1: The covariates included in ATRIA, Framingham, CHADS$_2$, CHA$_2$DS$_2$VASc and Q-Stroke

Predictor	ATRIA	Framingham	CHADS2	CHA2DS2VASc	Q-Stroke
Age 85+	x				
Age 75-84	x				
Age 65-74	x			x	
Age 60-62		x			
Age 63-66		x			
Age 67-71		x			
Age 72-74		x			
Age 75-77		x			
Age 78-81		x			
Age 82-85		x			
Age 86-90		x			
Age 91-93		x			
Age >93		x			
Age 75+			x	x	
Female	x	x		x	
Diabetes	x	x	x	x	x
Congestive heart failure	x		x		x
Prior Stroke or transient ischemic attack		x	x	x	
Hypertension	x		x	x	x
Systolic blood pressure$^+$		x			x
Total cholesterol: HDL* cholesterol ratio$^+$					x
Townsend deprivation score$^+$					x
Proteinuria	x				
eGFR*<45 or End stage renal disease	x				
Vascular disease				x	
Congestive heart failure or Liver disease				x	
Smoking status$^+$					x
Ethnicity$^+$					x
Coronary heart disease					x
Family history of congestive heart failure$^+$					x
Atrial fibrillation					x
Rheumatoid arthritis					x
Chronic renal disease					x
Valvular heart disease					x

Existing models for predicting stroke risk. + indicates predictors are often poorly recorded or missing in claims data.

* HDL - high-density lipoproteins, eGFR- estimated glomerular filtration rate.

Table 2: The internal and external validation performances of the existing stroke prediction models

	ATRIA	Framingham	CHADS2	CHA2DS2VASc	Q-Stroke
Internal c-statistic	0.72	0.66	0.82	0.61	0.65
External c-statistic					
UK Electronic Medical Records (EMR) 2015 [11]	0.7 (0.69-0.71)	-	0.68 (0.67-0.69)	0.68 (0.67-0.69)	-
Swedish EMR 2016 [12]	0.71 (0.70-0.71)	-	0.69 (0.69-0.70)	0.69 (0.69-0.70)	-
Taiwan 2016 [13]	-	-	0.66	0.70	-
New Zealand, Russia and the Netherlands 2014 [14]	-	0.70 (0.68-0.73)	-	-	0.71 (0.69-0.73)
UK EMR 2010 [15]	-	0.65 (0.63-0.68)	0.66 (0.64-0.68)	0.67 (0.65-0.69)	-

Internal and previously published external model fit statistics for each of the five models that predict stroke in atrial fibrillation patients

and outpatient pharmacy) from private healthcare coverage to employees, their spouses, and dependents, so patients are aged 65 or younger. The database contains data collected between 2000-2018.

IBM MarketScan® Medicare Supplemental Database (MDCR) represents health services of retirees in the United States with primary or Medicare supplemental coverage through privately insured fee-for-service, point-of-service, or capitated health plans. The patients are aged 65 or older. The database contains data collected between 2000-2018.

IBM MarketScan® Multi-State Medicaid Database (MDCD) contains adjudicated US health insurance claims for Medicaid enrollees from multiple states and includes hospital discharge diagnoses, outpatient diagnoses and procedures, and outpatient pharmacy claims as well as ethnicity. The database contains data collected between 2006-2018.

Optum© De-Identified Clinformatics® Data Mart Database – Socio-Economic Status (Optum Claims) is an adjudicated administrative health claims database for members with private health insurance. The population is primarily representative of US commercial claims patients (0-65 years old) with some Medicare (65+ years old) however ages are capped at 90 years. The database contains data collected between 2000-2018.

Optum© de-identified Electronic Health Record Dataset (Optum EHR) is a US electron health record containing clinical information, inclusive of prescriptions as prescribed and administered, lab results, vital signs, body measurements, diagnoses, procedures, and information derived from clinical Notes using Natural Language Processing (NLP). The database contains data collected between 2006-2018.

Stanford Translational Research Integrated Database Environment (STRIDE) is a clinical data warehouse that supports clinical and translational research at Stanford University. This resource includes the EHR data of approximately 2 million adult and pediatric patients cared for at either the Stanford Hospital or the Lucile Packard Children's hospital. This study was completed on an OMOP-CDM adherent instance of STRIDE. The database contains data collected between 2000-2018.

Columbia University Medical Center's (CUMC) data come from New York Presbyterian hospital's clinical data warehouse. The database comprises EHR data on approximately 5 million patients and includes information such as diagnoses, procedures, lab measurements and prescriptions. The database contains data collected between 1980-2018.

Ajou University School Of Medicine (AUSOM) is a database containing the entire EHR data from 1994 to 2018 of Korean tertiary hospital, Ajou university hospital. It contains medical record of about 2.9 million patients. The database contains data collected between 1994-2018.

The Integrated Primary Care Information (IPCI) is an electronic health care database containing patients of Dutch general practitioners (primary care). The database contains data collected between 1996-2018.

Each site had institutional review board approval for the analysis, or used deidentified data and thus the analysis was determined not to be human subjects research and informed consent was not deemed necessary at any site.

Participants

The existing models were applied to two target populations. Both target populations consisted of female patients newly diagnosed with atrial fibrillation and no prior stroke or anticoagulant use but target population 1 was patients aged 65 to 95 and target population 2 was all ages.

Target population 1: The target populations was defined as females aged 65-95 with either:

- 2 atrial fibrillation records
- 1 atrial fibrillation in an inpatient setting
- 1 atrial fibrillation with an electrocardiogram (ECG) within 30 days prior

and at least 730 days prior database observation and no prior stroke and no prior anticoagulant.

Target population 2: The target populations was defined as females with either:

- 2 atrial fibrillation records
- 1 atrial fibrillation in an inpatient setting
- 1 atrial fibrillation with an ECG within 30 days prior

and at least 730 days prior database observation and no prior stroke and no prior anticoagulant.

The target populations may contain different types of patients per database (e.g., different country US, European or Asian patients and different types of records such as inpatient and outpatient). The different databases used in this study are detailed in section 'Sources of data'.

Outcome

We predicted stroke occurring 1 day until 365 days after the initial atrial fibrillation start date. The stroke outcome was defined as:

- An ischemic or hemorrhagic stroke recorded with an inpatient or ER visit

The code sets used to define atrial fibrillation, ECG and ischemic or hemorrhagic stroke are presented in online Appendix B. The full analysis code (data creation and model evaluation) is available at: https://github.com/OHDSI/StudyProtocolSandbox/tree/master/ExistingStrokeRiskExternal Validation

Sensitivity analysis

Patients with a high risk of future stroke are often given anticoagulants as a preventative. If a high-risk patient is given an anticoagulant intervention during the 1-year time-at-risk this may prevent the stroke. We therefore performed a sensitivity analysis to remove patients who had an anticoagulant during the 1-year time-at-risk that may have prevented a stroke. For the sensitivity analysis, the target populations were modified by censoring patients at the point an anticoagulant was recorded, so any patient with an anticoagulant during the time-at-risk period was effectively removed from the target population unless they had a stroke prior to the anticoagulant.

Predictors

We calculated existing model predictors using phenotype definitions specified in the paper describing the development of the model when provided. If the development paper did not provide a definition, we used our own. The definitions for each predictor can be found in online Appendix A.

Missing Data

Age and gender are required by the OMOP common data model used by OHDSI and will never be missing.

For each condition (diabetes, chronic heart failure, stroke, hypertension, proteinuria, end stage renal disease (ESRD), vascular disease, liver disease, coronary heart disease (CHD), atrial fibrillation, rheumatoid arthritis, chronic renal disease and valvular heart disease), we considered no records of the condition in the database to mean the patient does not have the condition. Ethnicity is often missing completely from a database and when missing we did not include it. Smoking status and family history are rarely recorded in claims data, we imputed 0 (never smoker and no family history) when the predictor was missing. Townsend deprivation score is specific to the UK and was not included as a predictor in our validation. The blood pressure and cholesterol measurements are rarely recorded in claims data and were not included as predictors in our validation.

Statistical analysis

The prediction model performances were evaluated using the area under the receiver operating characteristic (AUROC) curve which is equivalent to the c-statistic for binary classification. Confidence intervals were also calculated when the number of outcome patients was fewer than 1000. As the models are being used to predict 1-year risk in diverse patients we recalibrated the models for each database. The models were recalibrated by fitting a linear model to the predicted scores to learn a database specific intercept and gradient. We present the calibration plots for each of the five models recalibrated in each of the datasets. For each decile we calculate the mean recalibrated predicted risk and plot against the observed fraction of patients who have the outcome.

Development vs Validation

We picked participants that matched all eligibility criteria for all 5 existing models being validated but this may be a subset of the patient population used to develop the model for many of the models. Many of the predictors for the Q-stroke model were not available in our data and the measurements for Framingham were also no available. The outcome in this validation study was 1 year following index but many of the models were developed for 10-year risk.

RESULTS

Participants

The characteristics of the participants across the network showed that hypertension was very common in the patients. Patients were older and often has renal and cardiac issues. See online Appendix C for the full characteristic table.

IPCI did not contain inpatient stroke records, so the models were unable to be evaluated on this dataset. The percentage of patients who had stroke recorded within 1 year in each of the remaining dataset target populations is presented in Table 3. The percentage of patients with stroke during the 1 year following atrial fibrillation diagnosis in the various target populations ranged from approximately 1% in CCAE, STRIDE, AUSOM and Optum EHR to 5% in MDCD and CUMC.

Model Performance

The results of the discriminative ability of the five existing models across all eight datasets that had inpatient stroke recorded are presented in Table 4. As the AUSOM and STRIDE datasets had outcome counts less than 100, we report the performance in Table 4 but do not include it in the aggregate summaries due to uncertainty in the estimates as a result of small sample sizes.

Across the datasets with sufficient outcome counts, ATRIA obtained a mean AUROC of 0.61 (range 0.57-0.64) on the female patients aged 65 or older and a mean AUROC of 0.63

Table 3: The stroke rate (% of target population) across the datasets

	Outcome rate % (Target population size)							
Target Population	CCAE	MDCD	MDCR	Optum claims	Optum EHR	CUMC	AUSOM	STRIDE
T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants	-	4.95 (25,880)	4.40 (89,156)	4.07 (110,905)	1.30 (149,906)	5.75 (4,312)	2.61 (268)	1.37 (3,366)
T2: Females with atrial fibrillation no prior stroke or anticoagulants	1.33 (61,224)	4.61 (33,262)	-	3.49 (139,376)	1.13 (189,815)	5.00 (5,758)	1.76 (455)	1.28 (4,456)
Sensitivity T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants (no anticoagulants during tar)	-	5.04 (23,586)	5.26 (56,511)	4.48 (78,353)	1.44 (99,212)	6.23 (3,403)	4.17 (144)	1.29 (2,094)
Sensitivity T2: Females with atrial fibrillation no prior stroke or anticoagulants (no anticoagulants during tar)	1.28 (46,054)	4.69 (29,546)	-	3.73 (100,757)	1.22 (128,409)	5.35 (4,546)	2.73 (256)	1.22 (2,786)

(Target population size in each dataset and the percentage of patients with stroke within 1 year of initial atrial fibrillation diagnosis)

(range 0.58-0.66) on the female patients of all ages. CHADS$_2$ obtained a mean AUROC of 0.58 (range 0.54-0.60) on the female patients aged 65 or older and a mean AUROC of 0.61 (range 0.56-0.63) on the female patients of all ages. CHA$_2$DS$_2$VASc obtained a mean AUROC of 0.60 (range 0.55-0.62) on the female patients aged 65 or older and a mean AUROC of 0.63 (range 0.58-0.65) on the female patients of all ages. Framingham obtained a mean AUROC of 0.60 (range 0.56-0.63) on the female patients aged 65 or older and a mean AUROC of 0.64 (range 0.57-0.65) on the female patients of all ages. Q-Stroke obtained a mean AUROC of 0.55 (range 0.53-0.56) on the female patients aged 65 or older and a mean AUROC of 0.57 (range 0.54-0.61) on the female patients of all ages.

The calibration plots showed that recalibrating the total scores using a linear model appears to work for ATRIA, Q-stroke, CHADS$_2$ and CHA$_2$DS$_2$VASc but the Framingham model may need a non-linear recalibration as it appeared to under-estimate risk in the middle risk groups, see online Appendix D.

Table 4: Discrimination performance of the existing models externally validated across the OHDSI datasets

		Database AUROC (95% CIs)							
Target Population*	Model	CCAE	MDCD	MDCR	Optum claims	Optum EHR	CUMC	AUSOM	STRIDE
T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants	ATRIA	-	0.57 (0.55-0.58)	0.63 (0.62-0.64)	0.61	0.62	0.64 (0.61-0.68)	0.60 (0.33-0.87)	0.49 (0.40-0.58)
	CHADS$_2$	-	0.54 (0.53-0.56)	0.60 (0.59-0.61)	0.59	0.60	0.60 (0.57-0.64)	0.51 (0.27-0.75)	0.48 (0.39-0.57)
	CHA$_2$DS$_2$VASc	-	0.55 (0.53-0.57)	0.60 (0.59-0.61)	0.59	0.62	0.61 (0.58-0.65)	0.53 (0.32-0.74)	0.52 (0.42-0.62)
	Framingham	-	0.56 (0.54-0.57)	0.62 (0.61-0.63)	0.59	0.61	0.63 (0.60-0.66)	0.58 (0.33-0.83)	0.61 (0.52-0.70)
	Q-Stroke	-	0.53 (0.52-0.55)	0.56 (0.55-0.57)	0.55	0.56	0.55 (0.51-0.59)	0.56 (0.29-0.84)	0.50 (0.41-0.59)
T2: Females with atrial fibrillation no prior stroke or anticoagulants	ATRIA	0.62 (0.60-0.64)	0.58 (0.56-0.59)	-	0.65	0.65	0.66 (0.62-0.69)	0.73 (0.58-0.89)	0.52 (0.44-0.60)
	CHADS$_2$	0.61 (0.59-0.62)	0.56 (0.55-0.57)	-	0.62	0.63	0.63 (0.60-0.66)	0.63 (0.43-0.83)	0.50 (0.42-0.57)
	CHA$_2$DS$_2$VASc	0.63 (0.61-0.65)	0.58 (0.56-0.59)	-	0.64	0.65	0.64 (0.61-0.67)	0.73 (0.60-0.85)	0.55 (0.47-0.62)
	Framingham	0.62 (0.60-0.64)	0.57 (0.56-0.59)	-	0.64	0.65	0.65 (0.62-0.68)	0.70 (0.53-0.86)	0.61 (0.53-0.69)
	Q-Stroke	0.61 (0.59-0.63)	0.54 (0.53-0.56)	-	0.57	0.58	0.56 (0.53-0.60)	0.63 (0.39-0.88)	0.51 (0.43-0.59)
Sensitivity T1: Females aged 65+ with atrial fibrillation no prior stroke or anticoagulants (no anti-coagulants during 1 year time-at-risk)	ATRIA	-	0.56 (0.55-0.58)	0.63 (0.62 -0.64)	0.61 (0.61 -0.62)	0.63 (0.61 -0.64)	0.65 (0.62-0.69)	0.69 (0.43-0.95)	0.55 (0.47-0.62)
	CHADS$_2$	-	0.54 (0.53-0.56)	0.61 (0.60 -0.62)	0.59 (0.58-0.60)	0.61 (0.59 -0.62)	0.62 (0.58-0.65)	0.61 (0.36-0.85)	0.51 (0.38-0.63)
	CHA$_2$DS$_2$VASc	-	0.55 (0.54-0.57)	0.61 (0.60 -0.62)	0.59 (0.58-0.60)	0.63 (0.61 -0.64)	0.63 (0.59-0.66)	0.64 (0.45-0.83)	0.55 (0.42-0.67)
	Framingham	-	0.55 (0.54-0.57)	0.62 (0.61 -0.63)	0.59 (0.59-0.60)	0.62 (0.61 -0.63)	0.64 (0.61-0.68)	0.68 (0.44-0.93)	0.64 (0.53-0.74)
	Q-Stroke	-	0.53 (0.52-0.55)	0.57 (0.55-0.58)	0.55 (0.54-0.56)	0.57 (0.55 -0.58)	0.56 (0.52-0.60)	0.61 (0.30-0.92)	0.47 (0.35-0.58)

Table 4: Discrimination performance of the existing models externally validated across the OHDSI datasets *(continued)*

		Database AUROC (95% CIs)							
Target Population[*]	Model	CCAE	MDCD	MDCR	Optum claims	Optum EHR	CUMC	AUSOM	STRIDE
Sensitivity T2: Females with atrial fibrillation no prior stroke or anticoagulants (no anti-coagulants during 1-year time-at-risk)	ATRIA	0.63 (0.61 -0.66)	0.58 (0.56 -0.59)	-	0.67	0.67	0.67 (0.64-0.70)	0.79 (0.63-0.94)	0.53 (0.43-0.63)
	$CHADS_2$	0.62 (0.60 -0.65)	0.56 (0.55 -0.58)	-	0.64	0.65	0.64 (0.61-0.68)	0.72 (0.53-0.91)	0.51 (0.41-0.62)
	CHA_2DS_2VASc	0.65 (0.62 -0.67)	0.58 (0.56 -0.59)	-	0.65	0.67	0.66 (0.63-0.69)	0.81 (0.71-0.90)	0.55 (0.44-0.65)
	Framingham	0.64 (0.61 -0.66)	0.57 (0.56 -0.59)	-	0.65	0.66	0.66 (0.63-0.69)	0.76 (0.59-0.93)	0.62 (0.51-0.72)
	Q-Stroke	0.62 (0.60 -0.64)	0.55 (0.53 -0.56)	-	0.58	0.6	0.57 (0.53-0.61)	0.68 (0.42-0.94)	0.47 (0.36-0.57)

Discrimination performance of the existing models across the datasets. The AUROC 95% confidence intervals were only calculated when the outcome count was less than 1000. *- See section 'Participants' for full inclusion/exclusion criteria

DISCUSSION

This study demonstrated the ability to perform external validation across five different data sites with access to nine databases in a short period of time. The countries corresponding to each database spanned across the USA, Europe and Asia. This shows the OHDSI network and tools can be used by researchers to efficiently perform external validation of models developed using observational healthcare data. The datasets used for validating the existing models that predict stroke in female patients with atrial fibrillation had varied outcome rates (1%-6%) indicating differences between the data. Despite the differences between the datasets there was consistently moderate discriminative performance across the databases.

Interpretation

Excluding patients with an anticoagulant after atrial fibrillation who did not have a prior stroke increased the incidence rate for all databases except CCAE and STRIDE. This suggests many people under 65 who have a stroke within a year of initial atrial fibrillation diagnosis had a prior anticoagulant. This may be a consequence of different treatment of patients with atrial fibrillation who are under 65 compared to being 65 and older. Atrial fibrillation patients who are given an anticoagulant when they are younger than 65 may have other risk factors prompting the use of an anticoagulant.

The sensitivity analysis shows the AUROC performance of models when removing patients with an anticoagulant and no stroke or an anticoagulant prior to stroke is comparable or better, see Table 3. This makes sense, for example consider the hypothetical situation where a clinical risk model correctly assigns a high risk to a patient who will have a stroke, but this high risk leads to a clinician giving the patient anticoagulants before the stroke that prevent the stroke occurring. In this situation the model's performance will be negatively impacted because of the intervention as the model was correct to assign a high risk but was wrong due to the intervention preventing the stroke. This raises the issue of how to fairly evaluate models that are already being used clinically or in situations where existing guidelines are used to identify patients who should being given preventative medicine. A fair evaluation is simple when there is no clinical intervention, but complex when preventative medicine exists for the outcome.

The validation performance of the models replicated using the OHDSI patient-level prediction framework and validated across the OHDSI network are comparable with other published results. The Q-Stroke model performed the worst out of all the existing models, but this is likely due to many variables of that model being specific to the UK or are things that are missing from claims data (such as family history, smoking status and recent measurements). This may indicate that Q-Stroke is not transportable to the US population. In addition, the performances of the models were worse when applied to older females as age is a key predictor in many of the models. In future work it would be interesting to investigate applying more complex machine learning methods with data-driven predictor selection to learn more advanced models for predicting stroke in older patients with atrial fibrillation and no prior stroke.

Implications

The external validation was performed over 60 days by five different research sites. Utilizing the OHDSI collaboration to validate a new prognostic model would enable extensive external validation across diverse patient populations. In addition, this could be accomplished in significantly less time than the current process for external validation that takes more than three years on average for one other researcher to implement the model [4]. The large-scale external validation was only possible because i) the OMOP common data model and OHDSI standardizations enable sharing of analysis code and ii) collaboration that is possible due to the OHDSI network. We recommend researchers who develop prediction models gain insight into their model's transportability by utilizing the OHDSI network's external validation ability. All that is required is to replicate their models using the OHDSI Patient-level prediction framework, which would also enable other researchers to readily implement the model.

Limitations

The main limitation of this study was the correct replication of existing models. The reason external validation rarely occurs is that many published models lack certain details such as how to define variables, as code lists are often not published. As a best practice patient-level prediction models should provide full definitions for all variables in the model and provide the model. We used the model's variable definitions when published, but when these were not available, we used our own code sets to define the variables. Another limitation in this study is the limited target populations investigated. We chose females aged 65 or older with no prior stroke as that was the intersection of criteria used when developing the five existing stroke models but we also wanted to see the impact of restricting to older patients (as many models use age as a variable), so we included a second target population of all females with no prior stroke. In future work it would be interesting to investigate the performances of the models across many different target populations. Finally, although OHDSI contains a large network of databases, it may not be possible to validate every prediction model on each of the databases within the network. For example, some databases may not contain the criteria used to identify the target population (e.g., if the target population required a specific measurement), may not have certain predictors recorded or may not have the outcome recorded (e.g., if the outcome requires an inpatient record but the data only contain outpatient records). The databases may also have insufficient observation time (e.g., a model predicting 10-year risk of stroke may not be suitably evaluated in US claims data such as Optum claims where only 13% of patients have 5+ years of observation). Future works needs to be done to investigate how to interpret the results of external validation across heterogeneous datasets.

CONCLUSION

In this paper we demonstrated the ability to scale-up external validation by using a collaborative network where researchers share a common data structure. The existing prediction models were validated on 9 databases across 5 sites within two months. We recommend that researchers utilize the OHDSI network to externally validate their models at scale across multiple datasets to gain insight into the generalizability and/or transportability of their models.

In addition, the results show that the existing stroke in atrial fibrillation models do not perform well at predicting stroke in the target population of older females in datasets we investigated. This prompts further research into whether a better model can be developed.

Supplementary Information

The online version contains supplementary material available at https://rdcu.be/digCJ

Declarations

Ethics approval and concept to participate

All patient data included in this study were deidentified.

The New England Institutional Review Board determined that studies conducted in Optum, IBM CCAE, IBM MDCR, and IBM MDCD are exempt from study-specific IRB review, as these studies do not qualify as human subjects research.

CUMC, STRIDE and AUSOM had institutional review board approval for the analysis, or used deidentified data, and thus the analysis was determined not to be human subjects research and informed consent was not deemed necessary at any site.

Consent for publication

Not applicable

Availability of data and materials

The Optum, IBM CCAE, IBM MDCR, and IBM MDCD 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-research-databases) and Optum (contact at: http://www.optum.com/solutions/data-analytics/data/real-world-data-analytics-a-cpl/claims-data.html) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available.

Due to ethical concerns, supporting data cannot be made openly available for the CUMC, STRIDE and AUSOM datasets.

Competing interests statement

Jenna Reps is an employee of Janssen Research & Development and shareholder of Johnson & Johnson. Patrick Ryan is an employee of Janssen Research & Development and shareholder of Johnson & Johnson. Peter Rijnbeek works for a research group who received unconditional research grants from Boehringer-Ingelheim, GSK, Janssen Research & Development, Novartis, Pfizer, Yamanouchi, Servier. None of these grants result in a conflict of interest to the content of this paper.

Funding statement

This work was supported by a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea [grant number: HI16C0992]. This project has received funding from 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.

Contributorship statement

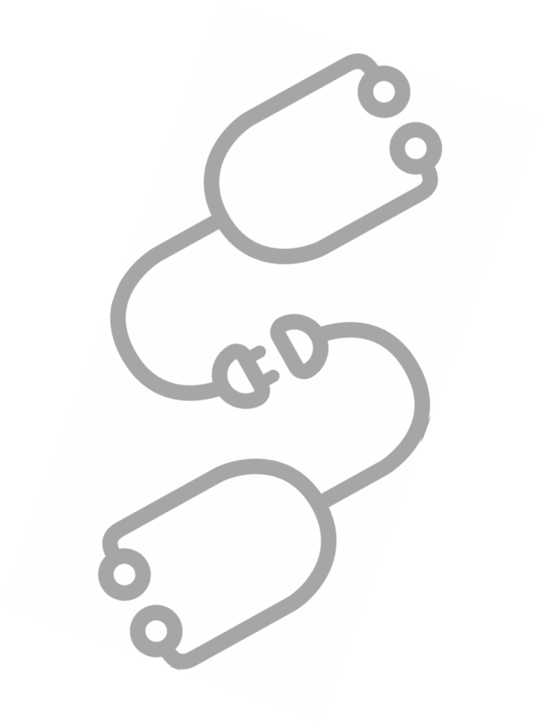
JMR lead and RDW, PBR, SCY, TF, EM, AC, RWP, HSL and PR contributed to the conception and design of the work, the analysis and the interpretation of data for the work. All authors contributed in drafting, revising and approving the final version.

Acknowledgements

Not applicable

BIBLIOGRAPHY

1. Obermeyer, Z. and Emanuel, E.J., Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*. 2016;*375*(13):1216.
2. Stewart, J., Manmathan, G. and Wilkinson, P., Primary prevention of cardiovascular disease: A review of contemporary guidance and literature. *JRSM cardiovascular disease*. 2017;*6*:1-9.
3. Van Calster, B., Wynants, L., Timmerman, D., Steyerberg, EW. and Collins, GS., Predictive analytics in health care: how can we know it works?. *Journal of the American Medical Informatics Association* 2019;*26*(12):1651-1654.
4. Siontis, GC., Tzoulaki, I., Castaldi, PJ, et al. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. *J Clin Epidemiol.* 2015;*68*(1):25-34.
5. Altman DG, Vergouwe Y, Royston P, et al. Prognosis and prognostic research: validating a prognostic model. *BMJ*. 2009;338:b605
6. Singer, DE., Chang, Y., Borowsky, LH., et al. A new risk scheme to predict ischemic stroke and other thromboembolism in atrial fibrillation: the ATRIA study stroke risk score. *J Am Heart Assoc.* 2013;*2*(3):e000250
7. Wang, TJ., Massaro, JM., Levy, D., et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community: the Framingham Heart Study. *JAMA*. 2003;*290*(8):1049-1056.
8. Gage, BF., Waterman, AD., Shannon, W., et al. Validation of clinical classification schemes for predicting stroke: results from the National Registry of Atrial Fibrillation. *JAMA*. 2001;*285*(22):2864-2870.
9. Lip, GY., Nieuwlaat, R., Pisters, R., et al. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. *Chest*, 2010;*137*(2):263-272.
10. Hippisley-Cox, J., Coupland, C. and Brindle, P. Derivation and validation of QStroke score for predicting risk of ischaemic stroke in primary care and comparison with other risk scores: a prospective open cohort study. *BMJ*. 2013;*346*:f2573.
11. van den Ham, HA., Klungel, OH., Singer, DE., et al. Comparative performance of ATRIA, CHADS2, and CHA2DS2-VASc risk scores predicting stroke in patients with atrial fibrillation: results from a national primary care database. *J Am Coll Cardiol*. 2015;*66*(17):1851-1859.
12. Aspberg, S., Chang, Y., Atterman, A., et al. Comparison of the ATRIA, CHADS2, and CHA2DS2-VASc stroke risk scores in predicting ischaemic stroke in a large Swedish cohort of patients with atrial fibrillation. *Eur Heart J.* 2016;*37*(42):3203-3210.
13. Chao, TF., Liu, CJ., Tuan, TC., et al. Comparisons of CHADS2 and CHA2DS2-VASc scores for stroke risk stratification in atrial fibrillation: which scoring system should be used for Asians?. *Heart Rhythm*. 2016;*13*(1):46-53.
14. Parmar, P., Krishnamurthi, R., Ikram, MA., et al. The Stroke Riskometer™ app: validation of a data collection tool and stroke risk predictor. *Int J Stroke*. 2015;*10*(2):231-244.
15. Van Staa, TP., Setakis, E., Di Tanna, GL, et al. A comparison of risk stratification schemes for stroke in 79 884 atrial fibrillation patients in general practice. *J Thromb Haemost*. 2011;*9*(1):39-48.



4

Using iterative pairwise external validation to contextualize prediction model performance: A use case predicting 1-year heart-failure risk in diabetes patients across five data sources

Ross D. Williams MSc[1], Jenna M. Reps PhD[2], Jan A Kors PhD[1], Patrick B Ryan PhD[2], Ewout Steyerberg PhD[3], Katia M. Verhamme MD[1], Peter R. Rijnbeek PhD[1]

[1] Department of Medical Informatics, Erasmus MC, University Medical Center Rotterdam, Rotterdam, The Netherlands.
[2] Janssen Research and Development, Titusville, NJ, USA.
[3] Department of Public Health, Erasmus MC, University Medical Center Rotterdam, Rotterdam

Link to article: https://rdcu.be/digDp

ABSTRACT

Introduction: External validation of prediction models is increasingly being seen as a minimum requirement for acceptance in clinical practice. The lack of interoperability of healthcare databases, however, has been the biggest barrier to this occurring at a large scale. Recent improvements in database interoperability enable a standardized analytical framework for model development and external validation. External validation of a model in a new database lacks context, whereby the external validation can be compared to a benchmark in this database. Iterative pairwise external validation (IPEV) is a framework which uses a rotating model development and validation approach to contextualize the assessment of performance across a network of databases. As a use case we predict 1-year risk of heart failure in patients with type 2 diabetes.
Methods: The method follows a 2-step process involving 1) development of baseline and data-driven models in each database according to best practices; 2) validation of these models across the remaining databases. We introduce a heatmap visualization that supports the assessment of the internal and external model performance in all available databases. As a use case, we developed and validated models to predict 1-year risk of heart failure in patients initializing a second pharmacological intervention for type 2 diabetes. We leveraged the power of the Observational Medical Outcomes Partnership Common Data Model to create an open-source software package to increase the consistency, speed and transparency of this process.
Results: A total of 403,187 patients were included in the study from 5 databases. We developed 5 models which when assessed internally had a discriminative performance ranging from 0.73 to 0.81 area under the receiver operating characteristic curve (AUC) with acceptable calibration. When externally validating these models in a new database, three models achieved consistent performance and in context often performed similarly to models developed in the database itself. The visualization of IPEV provided valuable insights. From this the model developed in the CCAE (Commercial Claims and Encounters) database is identified as the best performing model overall.
Conclusion"Using IPEV lends weight to the model development process. The rotation of development through multiple databases provides context to model assessment leading to improved understanding of transportability and generalizability. The inclusion of a baseline model in all modelling steps provides further context to the performance gains of increasing model complexity. The CCAE model was identified as a candidate for clinical use. The use case demonstrates that IPEV provides a huge opportunity in a new era of standardised data and analytics to improve insights and trust in prediction models at an unprecedented scale.

Key Points

1. External validation lacks context which inhibits understanding of model performance
2. Iterative Pairwise External Validation provides contextualised model performance across databases and across model complexity.

INTRODUCTION

External validation has been identified as an essential aspect of clinical prediction model development. It has previously been shown to be a key part of the evidence gathering process needed for creating impactful models that are adopted in the clinic (1). Currently, the majority of prediction models are not externally validated and where they are, they are poorly reported (2).

A major issue preventing the external validation of models is the lack of interoperability of healthcare databases (3). There are two main problems to solve. First, databases use different coding systems (e.g. International Classification of Diseases 10 (ICD-10) and SNOMED Clinical Terms), and second, the structure of these databases is different (4). A solution to this is to convert each database into a common format to improve syntactic interoperability and standardize to common vocabularies to improve the semantic interoperability.

After the format and vocabulary of these databases has been standardized it allows for the development of standardized tools and a framework for conducting prediction research (5, 6). Using these standard tools, and conducting research according to open science principles (7), removes many difficulties associated with externally validating prediction models. Some challenges remain, including the interpretation of results in the context of the new database. Furthermore, there are important privacy concerns that often need to be respected in the development process (8). For example, many data owners are unable to share patient-level data and as such any development process must be able to incorporate this (9).

Performance contextualization

Traditionally, a prediction model is trained on one database using predictors selected by domain experts and this model is then validated on other databases (10, 11). These models often consist of a limited number of predictors (12). Recently, data-driven approaches have been used to leverage all the information in the electronic health records which can result in models with many predictors. The question is how do we decide if the model works well on other databases? For this the standard approach is to compare the discriminative performance and model calibration with the performance obtained on the training data (13, 14, 15). If a performance drop is found then this could be because the model was tuned too much to the training data to properly transport to unseen data, i.e. the model was overfit or it needs recalibration. However, it could also be that the performance achieved is similar to the performance of a model that is trained on that same database. In other words, the model performs as good as possible in the context of the available data in that database. We need a model development approach that provides this context. Furthermore, simpler models are preferred as they are more easily clinically implemented and as such understanding the performance gain compared to the baseline of using only age and gender is valuable to contextualize the performance of the more complex model (16, 17).

In this paper we introduce Iterative Pairwise External Validation (IPEV), a framework to better contextualise the performance of prediction models, and demonstrate its value when developing and validating a prediction model in a network of databases. The use case for this model is to predict 1-year risk of heart failure following the initialisation of a secondary drug to treat T2DM. As described in detail in a literature review (18), the pathophysiological connection between diseases and their frequent adverse interactions should impact treatment choice (19). In the 2019 American Diabetes Association guidelines (20) it is advised to stratify patient treatments based upon established, or high risk of, heart failure (HF). Specifically, the guidelines state that thiazolidinediones (TZD) should be avoided in patients with heart failure and that in patients at high risk of heart failure Sodium-glucose co-transporter-2 inhibitors (SGLT2i) are preferred. The guidelines appear to be trending towards a more personalized treatment strategy (21, 22) and as such there is an opportunity to use risk prediction to further personalize treatment in the intermediate steps before treatment with insulin. This use case presents the opportunity to both evaluate IPEV and simultaneously create a potentially clinically impactful model.

METHODS

Analysis Methods

Iterative Pairwise External Validation

Iterative Pairwise External Validation (IPEV) is a new model development and validation procedure. It involves a 2-step procedure entailing, in the first step, creating two models per database, a model with only age and sex as covariates, which serves as a baseline for what a simple model can achieve, and a more complex data-driven model which assesses what the maximum achievable performance is. The second step is then validating these models both internally and also externally in the other databases. A diagram of this process can be seen in Figure 1.

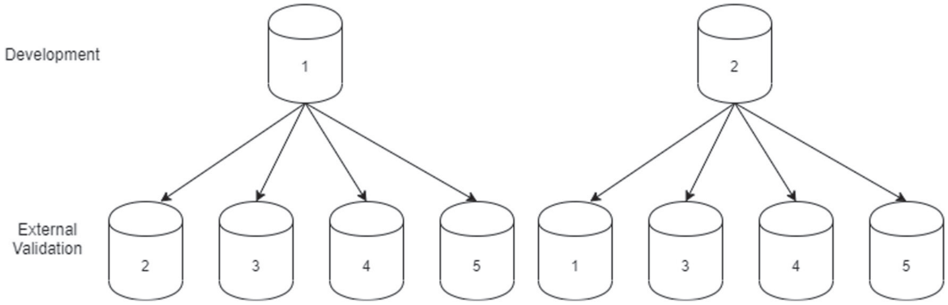


Figure 1 Rotation of databases for model development and external validation in the IPEV method.

Candidate covariates

Two sets of covariates are used to develop models. One set consists of only age and sex, and is used to create a baseline model. The other set is used to build a more complex data-driven model and consists of age, sex, and binary variables indicating the presence or absence of comorbidity (based on presence of disease codes) any time prior to index, and of procedures and drugs that occurred in the year prior to index date. The binary variables constructed are for any condition, procedure or drug that is in the history of the patient. For example, if any patient has a diagnosis of liver failure recorded in their medical records prior to the index date, then we create a candidate binary variable named 'liver failure any time prior' that has a value of 1 for patients with a record of liver failure in their history and 0 otherwise.

The use of these two sets of covariates shows the achievable performance for a simple set of covariates which can then be used to assess any added value of a more complex model. This gives a context to the performance gains relative to the increased model complexity.

Evaluation Analysis

For performance analysis we consider the area under the receiver operating characteristic curve (AUC) as a measure of discrimination. An AUC of 0.5 corresponds to a model randomly assigning risk and an AUC of 1 corresponds to a model that can perfectly rank patients in terms of risk (assigns higher risk to patients who will develop the outcome compared to those who will not). For calibration assessment we use calibration graphs and visually assess whether the calibration is deemed to be sufficient.

Proof of Concept

Predicting 1-year risk of developing heart failure (HF) following initiation of a second pharmaceutical treatment for type 2 diabetes mellitus (T2DM) was selected as a proof of concept. This case study could help inform treatment decisions by comparing an individual patient's risk of HF with the known safety profiles of the different medications.

Data Sources

The analyses were performed across a network of five observational healthcare databases. All databases contained either claims or EHR data from the US and have been transformed into the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM), version 5 (23). .

Table 1 describes the databases that are included in this study. The complete specification for the OMOP CDM, version 5 is available at https://ohdsi.github.io/CommonDataModel/cdm531.html.

Table 1 Database characteristics

Database	Acronym	Country	Data type	Time period	Database size (million patients)
Optum® de-identified Electronic Health Record Dataset	Optum EHR	US	EHR	2006-2018	87
IBM MarketScan® Commercial Database	CCAE	US	Claims	2000-2018	155
IBM MarketScan® Multi-State Medicaid Database	MDCD	US	Claims	2006-2017	30
IBM MarketScan® Medicare Supplemental Database	MDCR	US	Claims	2000-2018	10
Optum® De-Identified Clinformatics® Data Mart Database	Optum Clinformatics	US	Claims	2000-2018	98

Cohort definitions

Target Cohort

The target population consisted of T2DM patients who were treated with metformin and who became new adult users of one of Sulfonylureas, Thiazolidinediones, Dipeptidyl peptidase-4 inhibitors, Glucagon-like peptide-1 receptor agonists, or SGLT2is. The index date is the first prescription of one of these secondary treatments. We required all subjects to have a T2DM diagnosis, which was based upon the presence of a disease code and use of Metformin prior to the index date. Patients with HF or patients treated with insulin on or prior to the index date were excluded from the analysis. Patients were required to have been enrolled for at least 365 days before cohort entry.

Outcome definitions

The outcome was defined using the presence of a diagnosis code of HF occurring for the first time in the patient's history, between 1 and 365 days post index.

The cohort definition is available at: https://github.com/ohdsi-studies/PredictingHFinT2DM/tree/main/validation/inst/cohorts

The study period contained data from 2000-2018. The exact period varies between the databases and is available in Table 1.

Covariates

In total, we derived around 39,000 candidate covariates. These included more than 26,000 conditions, 13,000 procedures and drugs, and demographic information.

Statistical Analysis

Model development followed the framework for the creation and validation of patient-level prediction (PLP) models presented in Reps et al. (5). We used a 'train-test split' method to perform internal validation. In each target population cohort, a random sample of 75% of the patients (`training sample') was used to develop the prediction model and the remaining 25% of the patients (`test sample') was used to internally validate the prediction model developed.

We used regularized logistic regression risk models, also known as least absolute shrinkage and selection operator (LASSO). Regularisation is a process to limit overfitting in model development. This process works by assigning a "cost" to the inclusion of a variable and the variable must contribute more to the model performance than this cost in order to be included. If this condition is not met then the coefficient of the covariate becomes 0, which therefore eliminates the covariate from the model providing an in-built feature selection (24).

Open source software

We used the PatientLevelPrediction R-package (version 4.0.1) and R (v4.0.2) to perform all analyses. All development analysis code and cohort definitions are available at: https://github.com/ohdsi-studies/PredictingHFinT2DM

The validation package is available here: https://github.com/ohdsi-studies/PredictingHFinT2DM/tree/main/validation

RESULTS

Across all databases we selected 403,187 T2DM patients initiating second-line treatment. Of these, 12,173 developed HF during the one-year follow-up. Next, patient-level prediction of HF was performed. The number of patients and the AUCs are given in Table 2.

The AUC results, as shown in Figure 2, show reasonable performance. The main diagonal of the heatmaps show the internal validation. All other results are from external validation. The mean AUCs across internal and external validation were 0.78 (CCAE), 0.76 (MDCD), 0.76 MDCR, 0.78 (Optum Clinformatics), and 0.78 (Optum EHR). The best performing models in

Table 2 Number of patients and internal validation performance per database

Database	No. of T2DM patients	No. of HF patients	Incidence (%)	Age in years Mean (SD)	Female (%)	Full model AUC	Age Sex AUC
CCAE	112,989	1,843	1.6	53 (8)	46	0.78	0.64
MDCD	15,860	650	4.1	50 (12)	64	0.77	0.65
MDCR	22,433	1,658	7.4	73 (6)	48	0.73	0.64
Optum Clinformatics	92,272	4,332	4.7	63 (13)	48	0.80	0.69
Optum EHR	159,633	3,690	2.3	58 (12)	49	0.81	0.71

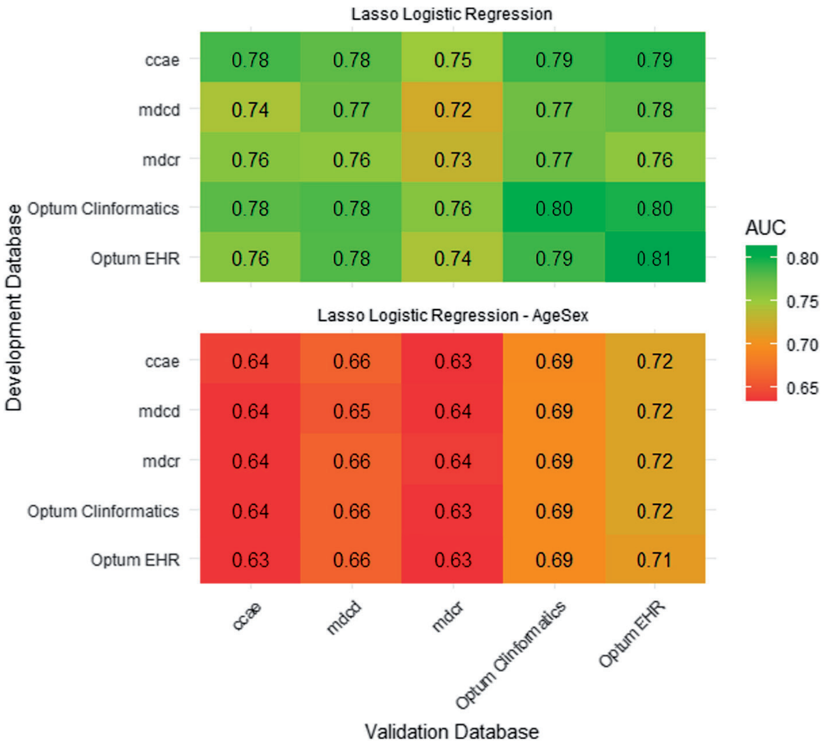


Figure 2 A heatmap of the AUC values across internal validation (values on the lead diagonal) and external validations of the developed prediction models. The colour scale runs form red (low discriminative ability) to green (high discriminative ability. The upper section details the performances for the data driven model. The lower half details the same but then for the Age and Sex model. Abbreviations: CCAE: Commercial Claims and Encounters, mdcd: Medicaid, mdcr: Medicare, optum EHR: optum electronic health records.

terms of discrimination were developed in CCAE, Optum ClinFormatics and Optum EHR and appear to be the most consistent across the external validations. When comparing the baseline model, consisting of only age and sex, with the full model the performances drops. For example, for CCAE the data-driven model achieves 0.78 compared to the baseline model of 0.64 and similarly for Optum Clinformatics with 0.80 (data-driven) and 0.69 (baseline).

Of note is that models externally validated in the MDCR dataset consistently outperformed the model that was developed there. This occurred for the data-driven model (internal: 0.73) with the external validation of CCAE, Optum Clinformatics and Optum EHR achieving 0.75, 0.76, 0.74 respectively.

We assessed the calibration of the three models with the best discrimination (CCAE, Optum Clinformatics and Optum EHR). The calibration results from these 3 models across the external validations are shown in Figure 3. The models generally appear to be well calibrated.

Concerning the best model produced, the CCAE and Optum Clinformics had the best discrimination performance. The CCAE model contained 195 covariates, compared to 413 for

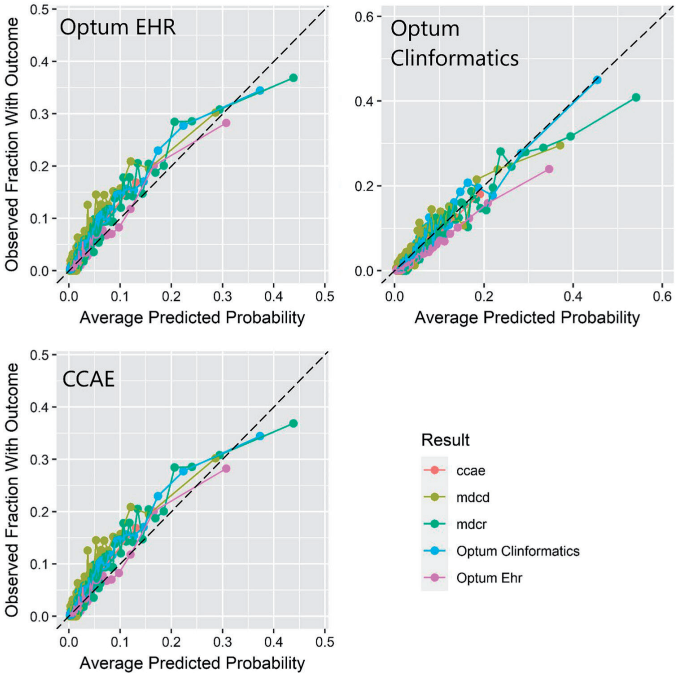


Figure 3 Internal and external calibration of the Optum EHR, Optum Clinformatics and CCAE trained models

Optum Clinformatics, and as such is preferred. The names and coefficients of the covariates in the CCAE model are available in online Appendix 1.

For the CCAE developed model demographic plots are provided in the electronic supplementary material. These plots show the calibration of the model stratified by sex across age groups.

All results are available in a study application located at: https://data.ohdsi.org/PredictingHFinT2DM/

DISCUSSION

This study demonstrates the use of IPEV for model development and external validation. External validation of a prediction model has traditionally lacked any contextual information on what the expected performance in the database should be. By including a baseline and data-driven model developed in each database, context can be added to the performance of a model externally validated in this database.

Due to the recent improvements in database interoperability and standardisation of tools, it was possible to utilise IPEV to develop and contextually validate models for predicting HF in

T2DM. This contextual validation provides a more rigorous approach to model assessment. For example, in the case where a model's performance drops from training to external validation but achieves performance consistent with expectations in the external validation database, this then raises the question of what the difference is between the two databases. Similarly, if a model achieves a lower performance than expected in a new database, then this can be interpreted as overfitting to training data.

The inclusion of a baseline model (using only age and sex covariates) in each training step provides context to the performance gain from increasing model complexity. By comparing the more complex model with this baseline model, a better assessment of complexity-performance trade-off can be made to analyse the potential for clinical implementation. If a large disparity in performance between these two models is observed then a parsimonious model (of around 10 variables) could be created to attempt to bridge the gap between the performance of the complex model and the ease of implementation of the baseline model. The interpretation of the results is aided by the inclusion of a heatmap. This allows for easy visual inspection of performance across external validations. Once differences in performance across external validation have been demonstrated, it would be interesting to investigate the case-mix of the cohorts in the database as well as the prevalence of the predictors to better understand these performance differences (25).

Considering the specific use case, the performance of the CCAE model developed in this paper suggests it could be used in treatment planning. This model has good discriminative performance that is consistent across external validations (AUC internal: 0.78, external 0.75-0.79). There is a minor loss in discrimination for some of the external validations, for example MDCR has the lowest AUC (0.75). This lower performance is in-line with the databases internal validation, and MDCR performs worst across all the external validations suggesting it is a more problematic dataset in which to make predictions. Possible explanations of this are that the underlying case-mix of patients could mean discrimination is harder. For example, patients in this database are generally older and as such it could become more difficult to separate them, there is also little to no overlap in ages of patients between CCAE and MDCR. Another reason could be the lower numbers of patients might mean there is insufficient data to provide a reliable estimate, or to develop the optimal model. Specifics of performance in different demographics is available in the shiny application. The model showed reasonable calibration across internal and external validations with some overestimation of risk for the higher risk patients. The Optum EHR external validation showed a larger miscalibration and could benefit from some recalibration before implementation. When comparing the data-driven models to the baseline models. The baseline models had only moderate performance across all the validations for all models, often there was a drop of between 0.1 and 0.2 AUC demonstrating that the increase in complexity provides significant performance gains. Age and sex alone are not sufficient to accurately predict future HF and more complex models are needed.

Calibration is important when using a model for clinical decision making and this result highlights that our model likely requires recalibration when applied to case-mixes that differ from the development database.

Considering the implementation of the model, this could occur either at a treatment facility or health authority level. Using the previously discussed ADA treatment guidelines, the use of a risk model to stratify patients can be impactful and the evidence generated in this paper suggests the CCAE developed model can be a candidate for clinical use. As patients can be assessed on their risk of HF, their treatment can be personalised helping to prevent medication switching or the addition of new medicines to treat HF when there are diabetes treatments with known beneficial HF effects. To our knowledge this is the only model that is available in open source that can be used for this specific prediction problem.

This method is scalable and can be expanded to use more databases as they are available. An example is through the EHDEN project, which is currently standardizing 100 databases to the OMOP CDM. This network could be leveraged to provide context to the external validation of prediction models at an unprecedented scale. This would lead to improved models, stronger evidence and a bigger clinical impact. When considering the case of a federated data network such as EHDEN, IPEV is particularly suitable. As privacy concerns prevent the sharing of patient-level data, a development and validation process that does not require this is necessary. IPEV incorporates "privacy by design" whereby, research can be performed by separate researchers at separate locations without the need to share patient data. This is a major advantage as it maintains the possibility to produce excellent and clinically impactful research without introducing any new privacy or security concerns. This means that the method can be used under the standard procedures of obtaining IRB approval, maintain the security of data and improve the quality of research, without significantly burdening the researchers.

A limitation of this method is that it does use the full data available for training. There is evidence to suggest that combining data can improve the internal validation. This however requires researchers to share data and violates data privacy concerns. Further, methods such as federated learning are compatible with IPEV. If a researcher is particularly concerned with improving the performance of the developed model they could combine n-1 databases and test in the nth. Then rotate through development using IPEV leaving out one database at a time. Increasing the data available for training and maintaining external validity simultaneously.

CONCLUSION

Using IPEV lends weight to the model development process. The rotation of development through multiple databases provides context allowing for thorough analysis of performance. The inclusion of a baseline model in all modelling steps provides further context to the performance gains of increasing model complexity. IPEV provides a huge opportunity in a new era of standardised data and analytics to improve insights and trust in prediction models at an unprecedented scale.

Supplementary Information

The online version contains supplementary material available at https://rdcu.be/digDp

Declarations

Funding

This project has received support from the European Health Data and Evidence Network (EHDEN) project. EHDEN received funding from 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.

Conflicts of interest/Competing interests

Ross D. Williams, Jan A. Kors and Ewout Steyerberg report no conflict of interest. Katia M. Verhamme and Peter R. Rijnbeek works for a research group that received unconditional research grants from Boehringer-Ingelheim, GSK, Janssen Research & Development, Novartis, Pfizer, Yamanouchi, Servier. Jenna M. Reps and Patrick B. Ryan are employees and shareholders of Janssen Research & Development and shareholder of Johnson & Johnson.

Availability of data and material

Ross D. Williams is responsible for the data. Due to privacy concerns the patient-level data for the study is not available. All results are available at: https://data.ohdsi.org/PredictingHFinT2DM/

Code availability

All code used in the study is provided open source at: https://github.com/ohdsi-studies/PredictingHFinT2DM

Authors' contributions

Ross D. Williams had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis Patrick B. Ryan and Katia M. Verhamme contributed significantly to the development of the cohort definitions. All contributed substantially to the study design, data analysis and interpretation, and the writing of the manuscript. All authors read and approved the final version.

Ethics approval

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

Consent to participate

The manuscript uses secondary data and as such no human participants were involved in the study and informed consent for participation was not necessary at any site.

Consent for publication

Not applicable.

BIBLIOGRAPHY

1. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016 Jan;69:245-7.
2. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014 Mar 19;14:40.
3. Lehne M, Sass J, Essenwanger A, Schepers J, Thun S. Why digital medicine depends on interoperability. NPJ Digit Med. 2019;2:79.
4. Kent S, Burn E, Dawoud D, Jonsson P, Ostby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2021 Mar;39(3):275-85.
5. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018 Apr 27.
6. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020 May 6;20(1):102.
7. Woelfle M, Olliaro P, Todd MH. Open science is a research accelerator. Nature chemistry. 2011;3(10):745-8.
8. Mooney SJ, Pejaver V. Big Data in Public Health: Terminology, Machine Learning, and Privacy. Annu Rev Public Health. 2018 Apr 1;39:95-112.
9. Mandl KD, Kohane IS. Federalist principles for healthcare data networks. Nat Biotechnol. 2015 Apr;33(4):360-3.
10. Moons KG, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: I. Development, internal validation, and assessing the incremental value of a new (bio) marker. Heart. 2012 May;98(9):683-90.
11. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): the TRIPOD Statement. Br J Surg. 2015 Feb;102(3):148-58.
12. 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 May 16;353:i2416.
13. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012 May;98(9):691-8.
14. 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 Jun 22;353:i3140.
15. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clinical Kidney Journal. 2020;14(1):49-58.
16. Helgeson C, Srikrishnan V, Keller K, Tuana N. Why Simpler Computer Simulation Models Can Be Epistemically Better for Informing Decisions. Philosophy of Science. 2021;88(2):213-33.
17. Zhang J, Wang Y, Molino P, Li L, Ebert DS. Manifold: A Model-Agnostic Framework for Interpretation and Diagnosis of Machine Learning Models. IEEE Trans Vis Comput Graph. 2019 Jan;25(1):364-73.
18. Tousoulis D, Oikonomou E, Siasos G, Stefanadis C. Diabetes Mellitus And Heart Failure. European Cardiology Review. 2014;9(1):37-42.
19. Nichols GA, Hillier TA, Erbey JR, Brown JB. Congestive heart failure in type 2 diabetes: prevalence, incidence, and risk factors. Diabetes Care. 2001 Sep;24(9):1614-9.

20. Care F. Standards of Medical Care in Diabetes 2019. Diabetes Care. 2019;42(Suppl 1):S124-S38.

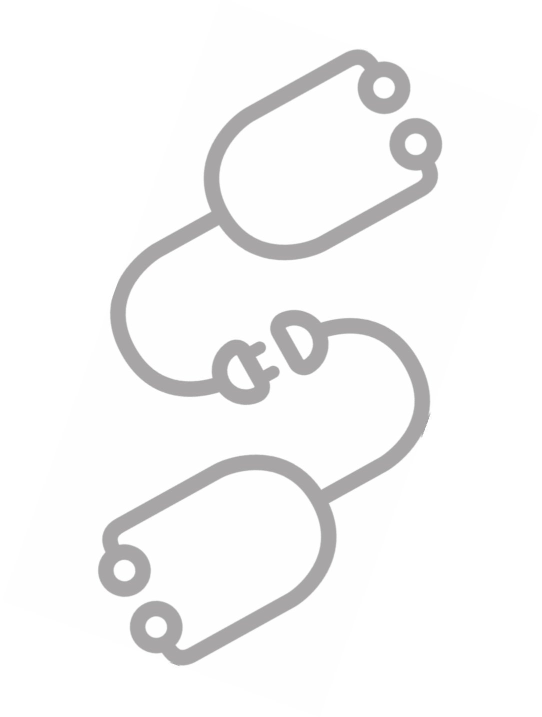
21. Association American D. Updates to the Standards of Medical Care in Diabetes-2018. Diabetes Care. 2018 Sep;41(9):2045-7.

22. Marathe PH, Gao HX, Close KL. American Diabetes Association Standards of Medical Care in Diabetes 2017. J Diabetes. 2017 Apr;9(4):320-4.

23. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012 Jan-Feb;19(1):54-60.

24. Tibshirani R. Regression shrinkage and selection via the Lasso. J Roy Stat Soc B Met. 1996;58(1):267-88.

25. 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 Mar;68(3):279-89.



5

The DELPHI library: Improving model validation and dissemination through a centralised library of prediction models

Ross D. Williams[1]
Sicco den Otter[1]
Luis Henrik John[1]
Jenna M. Reps[12]
Peter R. Rijnbeek[1]

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
[2]Janssen Research and Development, Titusville, NJ, USA

Article under submission

ABSTRACT

Background: Clinical prediction modelling has seen a rapid rise in interest in the last 10 years, but implementation of models in clinical practice still lags behind their development. This is mainly because the reporting on the development and performance of prediction models is sub-optimal, hampering reproducibility and extensive external validation and updating. It is difficult to reproduce the model development due to lack of data interoperability and standardisation of development steps. The Observational Health Data Sciences and Informatics (OHDSI) initiative has developed a framework for prediction model development and validation that enforces best practices. This paper introduces the DELPHI library, a database and a graphical user interface for sharing, finding, assessing, and validating clinical prediction models developed within the OHDSI framework. The aim is to follow the Findable, Accessible, Interoperable, and Re-usable (FAIR) principles to improve prediction modelling using observational data. Future validation and updating studies for prediction model are expected to benefit from this library.

Main body: A database structure was created to store all relevant information necessary for a fully reproducible model development process, and to improve transparency on model performance measures. This data is shared in a graphical user interface which allows independent researchers, clinicians, and regulators to access, explore, and assess the models on their own data. As a proof-of-concept study we describe how the DELPHI library was used to share 53 models and their performance. The library is publicly available and will expand as more models are developed, validated or updated under the OHDSI framework.

Conclusions: The OHDSI prediction framework in combination with the DELPHI library makes prediction models more FAIR. DELPHI enables reproducibility of model development and large-scale external validation. This is an important prerequisite for their clinical adoption.

BACKGROUND

Over the past decade there has been a rapid increase in the number of published clinical prediction models (1). There has not however been a similar rise in the use of these models within clinical practice. This gap is due to multiple reasons including insufficient reporting, e.g., models are not shared publicly (2), opaque development and validation methods, insufficient testing (3), creating a lack of trust from clinical stakeholders. Many models are also often developed without a clear clinical use case or implementation strategy in the clinical setting (4).

Recent efforts to improve the utility of clinical prediction models include the Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD) (2), which aimed to standardise and improve the reporting of models. Since the publishing of the TRIPOD guidelines there has been some improvement in the standards of reporting. However, many models still do not follow best practices in either model development, validation, or reporting (1). There is a clear need to further enforce the adherence to these best practices (5-7). Another effort to improve model dissemination is the Tufts Clinical Prediction Model repository (8). This provides a location for storing published clinical prediction models, in this article referred to as Patient-Level Prediction (PLP) models, with short recommendations for use of the models and often an attached scientific article detailing the development process. What this repository lacks, however, is the ability to explore the results of the model training and validation interactively, and it does not allow to download the model and execute it locally against data.

In order to create a repository with this functionality, we first need to improve the interoperability of the data as well as the prediction models. Differences in data structure (syntactic interoperability) and terminology (semantic interoperability) make it hard to enforce a standardised and reproducible development and validation process. When models are developed on databases without a common data model, each model will have to be transformed to the format of the database that it needs to be applied in. Furthermore, if a model is developed using a database that has diagnoses recorded using International Classification of Diseases (ICD-10) codes, the model will need to be translated to be applied in a database that is based on International Classification of Primary Care (ICPC) codes. This is not scalable to many prediction models and databases.

In an ideal situation a clinician or researchers interested in a specific prediction problem, can search for all relevant prediction models, assess the available performance measures, and download the model from a central repository to evaluate the model on its own data. The newly obtained model performance can then be added to this central repository to expand the body of knowledge. This creates an open science environment for prediction modelling that is dynamic and fully transparent.

Observational Medical Outcomes Partnership Common Data Model

Improving interoperability of data requires the use of a Common Data Model (CDM).The recent widespread adoption of common data models (such as the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) (9) many of the barriers have been removed. The OMOP CDM improves the syntactic and semantic interoperability of the data. For example, two different databases, database A and database B, have different database structures (e.g., tables and columns) and coding systems. If the two databases were mapped to the OMOP CDM, then database A OMOP CDM and database B OMOP CDM have the same structure and coding system. This means any data extraction code written for database A OMOP CDM can also be applied to database B OMOP CDM. By using the OMOP CDM it alleviates many of the burdens associated with data extraction for a particular study. When considered in the context of a prediction model, what it allows is for a standard set of tools to be developed to create, train and evaluate a prediction model based upon the known format of the data model. The feature extraction is done using the same code, exactly the same way, for any OMOP CDM database. However, for example, when a model is trained in a non-OMOP CDM database (diagnoses coded using ICD-10) and applied to a non-OMOP CDM database (diagnoses coded using ICPC), the researcher would have to manually write code to extract model features using the ICPC coding on a per study basis. This can mean a model changes when validated.

Patient-Level Prediction

A prediction task can be thought as trying to map a set of *predictive variables* (e.g., history of diabetes, age, sex at birth) to an outcome label (e.g., will develop cancer) in a dataset. When learning this mapping, or function, an algorithm attempts to learn a set of parameters for predictive variables to better predict the labelled outcome. Patient-Level prediction (PLP) has been extensively defined elsewhere (5-7), however in short a prediction task consists of a target cohort (those for whom we want to make a prediction), and outcome (the outcome of interest to be predicted) and a time at risk (during which the outcome is being predicted relative to the target cohort index). To develop a model for a given prediction task, the user needs a suitable dataset and must specify the modelling design (data pre-processing, type of classifier, hyper-parameter search). A standardised pipeline for the development and validation of models has been produced that enforces these guidelines and produces models with a flexible algorithm format and a standardised format. This pipeline is implemented using code from the PatientLevelPrediction R package.

The OHDSI PatientLevelPrediction R package software provides standardised tools for developing and validating prediction models using data in the OMOP CDM. The PatientLevelPrediction package takes standard inputs for the database, the prediction task and the modelling design and outputs a standardized structure containing the final model, the model design (including the prediction task), internal predictions, internal validation performance (AUC,AUPRC, calibration statistics, etc.) and meta data about the process. The package can be accessed at: https://github.com/OHDSI/PatientLevelPrediction.

Findable, Accessible, Interoperable and Re-usable (FAIR)

The FAIR guiding principles are intended to improve the infrastructure supporting the reuse of data(10). A major component of this is to make available data and metadata contributing to, or produced by, scientific research projects. Given the privacy concerns involved in research using observational data, the release of the underlying data used to create prediction models is impossible. The DELPHI library leverages the interoperability of the OMOP CDM to be able to release all relevant model aggregate performance data and metadata (including the model itself and relevant study artefacts such as cohort definitions) in a unique, persistent and accessible manner. FAIR principles are key in the motivation and construction of the DELPHI library.

CONSTRUCTION AND CONTENT

In this paper we propose a centralised repository for PLP models that enables users to explore model parameters and model performance in addition to the ability to download and apply the models to new data.

The implementation of this repository consists of two separate parts, 1. a database storing all the specifications of the prediction models (e.g. target and outcome cohorts etc.), the performances (both internal and external), and information about the researchers that developed the models; 2. A graphical user interface (GUI) to connect to and interact with this database in a user-friendly manner. This user interface has the functionality to upload results, explore results and to download a software package to enable the external validation of results against new data in the OMOP CDM format.

This can then be shared with the community by uploading the external validation results to the DELPHI library.

Database

Due to the standard framework for prediction model development and validation implemented within OHDSI, models developed have a standardised output. This means that the creation of a relational database containing this information is straightforward. The relational database (PostgreSql) consists of 30 tables, each corresponding to an element of the standard output from the PLP framework. An entity relation diagram of the database is given in Figure 1.. The entity relation diagram shows 6 different interconnected sections. These are models, model development settings, performance, database information, researchers, and diagnostics. In the model sector, the models themselves are stored and there is a table for recalibration to allow for model updating to be done and maintain a link to the original model. The model develop-

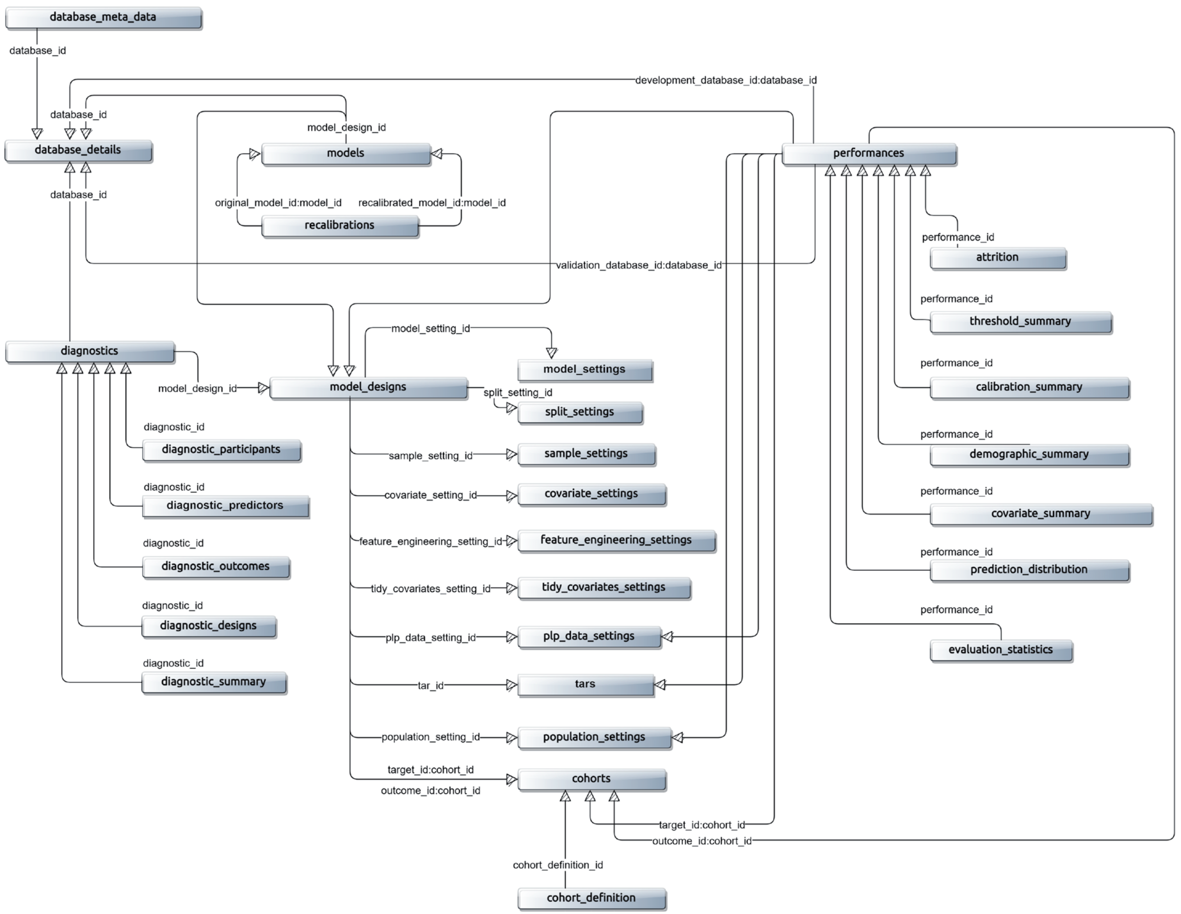


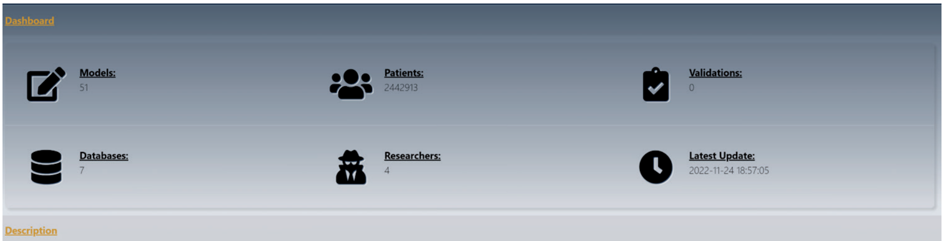
Figure 1 The entity-relation diagram detailing the structure and relationships between tables in the DELPHI database. The grey boxes contain the names of the tables and the tags on the arrows are the foreign keys

ment settings include all the information needed to replicate the development of the model. This is important for the reproducibility of the research. The performance contains all the aggregate performance statistics that are generated as standard output form the PLP R package. Importantly the prediction per patient any performance metrics generated with low counts (standard less than 5 patients) are removed before addition as this is considered patient specific information is therefore sensitive. The database information contains the specifics of which databases were used for development and validation and relevant meta data such as database type (general practice, claims etc). Researchers contains name and contact information for the researcher that performed the development or validation. The diagnostics section contains the results of the diagnostics that are performed on the analyses when using the PLP R package. This is a set of quality control measures that should aid in the generation of high-quality evidence.

Graphical User Interface

The processing of the data is done by the graphical user interface. The GUI was developed using .NET (core 3.1), a cross-platform, modular open-source software framework, and Entity Framework Core which is a modern object-database mapper built to integrate between SQL databases and the .NET framework. This allows for integration of our frontend with the PostgreSql database instance. The frontend was built in Typescript a version of Javascript that includes cross-browser, multi-platform support for large-scale applications. This combined with the React framework creates a UI that can adapt to multiple devices, views and platforms. The application is also supported by Electron which allows it to be run as a native windows or Mac application. The application is also supported by Docker. This way it is quick and easy to setup a new development environment or to install everything on a server, which is easy to maintain, update and reinstall.

The GUI has been developed to facilitate the uploading, searching and exploring of models and performance. The landing page of the library shown in Figure 2. This shows an overview of all the available models that have been developed and added to the system. The main library page displays all of the developed models along with high-level information including target and outcome cohorts, development database and some performance measures. This is detailed in



Models:
51

Patients:
2442913

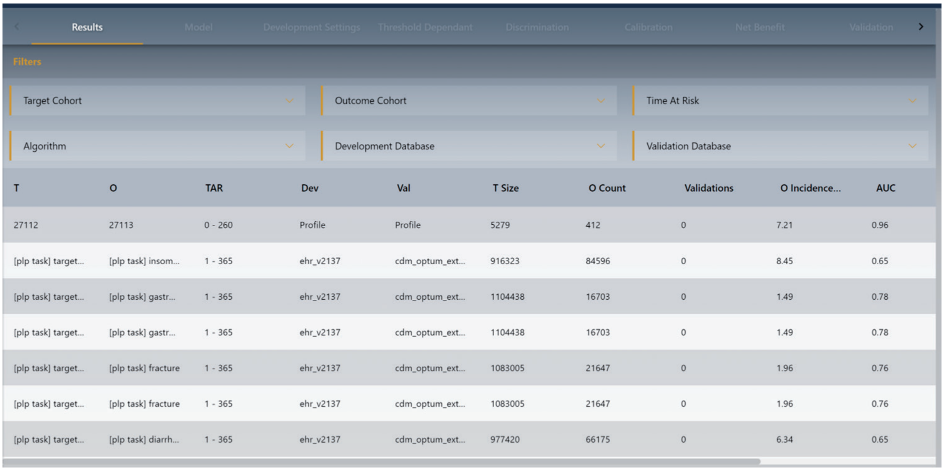
Validations:
0

Databases:
7

Researchers:
4

Latest Update:
2022-11-24 18:57:05

Figure 2 The landing page of the DELPHI library. This figure shows the number of models, patients, databases, researchers and external validations contained in the models in the databases.

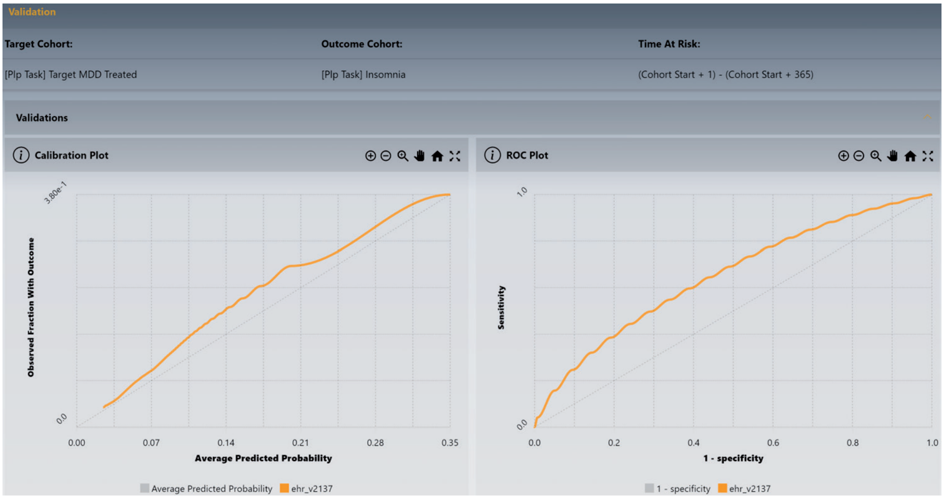


T	O	TAR	Dev	Val	T Size	O Count	Validations	O Incidence...	AUC
27112	27113	0 - 260	Profile	Profile	5279	412	0	7.21	0.96
[plp task] target...	[plp task] insom...	1 - 365	ehr_v2137	cdm_optum_ext...	916323	84596	0	8.45	0.65
[plp task] target...	[plp task] gastr...	1 - 365	ehr_v2137	cdm_optum_ext...	1104438	16703	0	1.49	0.78
[plp task] target...	[plp task] gastr...	1 - 365	ehr_v2137	cdm_optum_ext...	1104438	16703	0	1.49	0.78
[plp task] target...	[plp task] fracture	1 - 365	ehr_v2137	cdm_optum_ext...	1083005	21647	0	1.96	0.76
[plp task] target...	[plp task] fracture	1 - 365	ehr_v2137	cdm_optum_ext...	1083005	21647	0	1.96	0.76
[plp task] target...	[plp task] diarrh...	1 - 365	ehr_v2137	cdm_optum_ext...	977420	66175	0	6.34	0.65

Figure 3 The main repository page of the DELPHI library. This contains information on the target and outcome cohorts, the database used for the development and validation as well as performance metrics.

Figure 3. Once a model has been selected by the user various aspects of the results can be explored. This includes the discrimination and calibration performance as well as information about the selected model such as the specific settings used for development, model type and the database used. Furthermore, there is a section to explore external validation of the model. This is shown in figure 4. This displays comparisons of discrimination and calibration performance across the various OMOP CDM database that the model has been evaluated against.

Search functionality has been added to allow for researchers to search for different models using multiple fields. These include the target and outcome cohorts. For example, a user might



Target Cohort: [Plp Task] Target MDD Treated

Outcome Cohort: [Plp Task] Insomnia

Time At Risk: (Cohort Start + 1) - (Cohort Start + 365)

Figure 4 The validation tab detailing the ROC plot and calibration plot along with performance metrics to analyses the selected prediction model.

be interested in searching for prediction models for patients who will experience a stroke (an outcome cohort search) or they might be interested in modes that predict various outcomes for patients that have experience a major adverse cardiac event (a target cohort search). This is helpful for clinical researchers to access models relevant to their domain. They likely have a specific clinical question e.g., "What is the risk of stroke for a patient with newly diagnosed atrial fibrillation?". The GUI will provide them with search functionality to find the model that best matches this. Ideally there is a model for exactly this question, but they could also be interested in models that either match the target (e.g., a model predicting a non-stroke outcome for atrial fibrillation patients) or outcome (e.g. prediction of stroke in a not atrial fibrillation cohort). Once the relevant models have been identified, these models can be compared to see if one is fit for practice. The search result can also be filtered by researcher, by model type which again will allow clinical researchers to specify model types they deem to be acceptable for their practice and for methodological researchers to assess performance for varying model types across a multitude of problems.

There is also an option to download the model. This creates a JSON configuration file with the standardized model design settings that were used to develop the model plus the model (as a JSON, rds or python pickle file). The downloaded model can be used by the PatientLevelPrediction package to develop a new model with the same design in any OMOP CDM data or validate the model in any OMOP CDM data. The model validation process requires the user to download the model JSON from the repository, open an R session, load the JSON as an R object, enter their OMOP CDM connection details and then run the execute function. The results of this external validation can then be added back into the DELPHI database.

The current version of the database contains 53 models, from 5 different databases and using the data of more the 3.5m patients.

UTILITY AND DISCUSSION

The main purpose of prediction modelling is to aid decision making in clinical practice. In order to do this the level of trust in the modelling process and the models themselves needs to be improved. Currently prediction models are spread throughout the scientific literature, often the parameters of the models themselves are unavailable and the exploration of results is limited to what has been reported and it is often incomplete. This article details an application that revolutionises the field by centralising models and results and makes accessible everything that is needed to assess and implement prediction models. Models that are currently produced are often static, the article is published and little if any updates on evidence (new external validations) or model updates (recalibration) follow.

The creation of a database to store the standardised output of PLP models developed on the OMOP CDM increases the findability, accessibility, interoperability, and reusability of the models. Whereas previously an interested party would need to perform a literature search of disparate sources and keywords, they are now able to perform a systematic search within a single database to find available, ready for implementation models. Once a model is found, all relevant model information, for example the performance, definitions of target and outcome cohorts (especially relevant for clinical implementation) and importantly the covariates and model specification will be readily available.

As DELPHI develops, the set of PLP problem specifications developed by clinical researchers to answer relevant and impactful clinical questions will grow. This provides opportunities for the field to create of a set of FAIR benchmarking models for the testing of new algorithms and model development techniques. Currently, machine learning models are often developed and tested in relatively small, synthetically created benchmark datasets that do not capture the complexity of real-world data. Moreover, these lack a direct relationship to the questions in healthcare that clinical prediction models could help to answer. Within DELPHI, the more organically developed set of benchmark models can be used by methods researchers to improve their model development and evaluation techniques. This is a deviation from the traditional benchmarking conducted using identical datasets (11). Due to, amongst other things, patient privacy concerns and inherent biases (12), the provision of a benchmarking dataset from observational data is challenging (13). By instead providing a set of models, Benchmark models will receive a tag in the database for easy identification. This will provide a better, more relevant set of benchmarks and provide more relevant and impactful evidence on the performance of new techniques.

A limitation of this software is potentially its reliance on models developed against the OMOP CDM which precludes any models generated outside of this framework from being included. However, there are instructions available for translating existing model into the OHDSI standardized model format and if that is done the model can then be added to the DELPHI library repository. This flexibility is provided as well as technical support to map existing models to encourage researchers to submit models developed outside of the OMOP CDM to then receive easy and rapid validation of the models.

We believe that the DELPHI library represents a paradigm shift in the field of PLP modelling and as such should contribute to a vast improvement in the assessment and uptake of models in clinical practice.

CONCLUSION

The DELPHI library presents a significant improvement over the current situation in the field of PLP modelling. By creating a centralised standardised location for models and their performance the searchability and accessibility of PLP models is dramatically improved. This standardised format and accompanying software enforces best practices in model development and reporting and as such would help to raise the standard of clinical prediction modelling. It also moves the field away from a static publishing model to a dynamic ecosystem where models can continue to be analysed and updated on new data and in new settings. We believe that the improved flexibility in model exploration will aid in the adoption of PLP models in clinical practice.

Funding

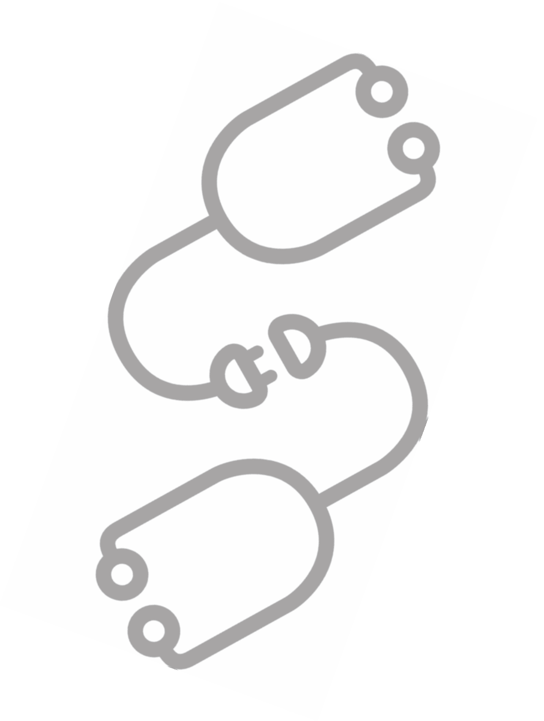
This project has received support from the European Health Data and Evidence Network (EHDEN) project. EHDEN received funding from the Innovative Medicines Initiative 2 Joint Undertaking (JU) under grant agreement No 806968. The JU receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

BIBLIOGRAPHY

1. Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022;29(5):983-9.
2. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
3. Steyerberg EW, Harrell FE, Jr. Prediction models need appropriate internal, internal-external, and external validation. J Clin Epidemiol. 2016;69:245-7.
4. Cave A, Brun NC, Sweeney F, Rasi G, Senderovitz T, Taskforce H-EJBD. Big Data - How to Realize the Promise. Clin Pharmacol Ther. 2020;107(4):753-61.
5. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
6. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018.
7. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernandez-Bertolin S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Comput Methods Programs Biomed. 2021;211:106394.
8. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. Diagn Progn Res. 2017;1:20.
9. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60.
10. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
11. Thiyagalingam J, Shankar M, Fox G, Hey T. Scientific machine learning benchmarks. Nat Rev Phys. 2022;4(6):413-20.
12. Denton E, Hanna A, Amironesei R, Smart A, Nicole H, Scheuerman MK. Bringing the people back in: Contesting benchmark machine learning datasets. arXiv preprint arXiv:200707399. 2020.
13. Paullada A, Raji ID, Bender EM, Denton E, Hanna A. Data and its (dis)contents: A survey of dataset development and use in machine learning research. Patterns (N Y). 2021;2(11):100336.

Part II

Clinical Applications



6

90-day all-cause mortality can be predicted following a total knee replacement: An international, network study to develop and validate a prediction model

Ross D. Williams MSc[1 ¶], Jenna M. Reps PhD[2 ¶],
The OHDSI/EHDEN Knee Arthroplasty Group,
Peter R. Rijnbeek PhD[1], Patrick B. Ryan PhD[2&],
Daniel Prieto-Alhambra PhD[3&*]

[1] Erasmus University Medical Centre, Rotterdam
[2] Janssen Research and Development, Raritan, NJ, USA
[3]NDORMS, University of Oxford, Oxford, UK

* Corresponding author
[¶]These authors contributed equally to this work.
[&]These authors contributed equally to this work.

Link to article: https://rdcu.be/digIZ

ABSTRACT

Purpose: The purpose of this study was to develop and validate a prediction model for 90-day mortality following a Total knee replacement (TKR). TKR is a safe and cost-effective surgical procedure for treating severe knee osteoarthritis (OA). Although complications following surgery are rare, prediction tools could help identify high-risk patients who could be targeted with preventative interventions. The aim was to develop and validate a simple model to help inform treatment choices.

Methods: A mortality prediction model for knee OA patients following TKR was developed and externally validated using a US claims database and a UK general practice database. The target population consisted of patients undergoing a primary TKR for knee OA, aged ≥40 years and registered for ≥1 year before surgery. LASSO logistic regression models were developed for post-operative (90-day) mortality. A second mortality model was developed with a reduced feature set to increase interpretability and usability.

Results: A total of 193,615 patients were included, with 40,950 in The Health Improvement Network (THIN) database and 152,665 in Optum. The full model predicting 90-day mortality yielded AUROC of 0.78 when trained in OPTUM and 0.70 when externally validated on THIN. The 12 variable model achieved internal AUROC of 0.77 and external AUROC of 0.71 in THIN.

Conclusions: A simple prediction model based on sex, age, and 10 comorbidities that can identify patients at high risk of short-term mortality following TKR was developed that demonstrated good, robust performance. The 12-feature mortality model is easily implemented and the performance suggests it could be used to inform evidence based shared decision-making prior to surgery and targeting prophylaxis for those at high risk.

INTRODUCTION

TKR surgery is generally a safe procedure with fewer than 10% of patients experiencing post-operative complications. These adverse events include short-term (e.g. 90-day) post-operative mortality (1, 2). Mortality following TKA is low and has been declining over recent years (3). However, there is a scarcity of data on who is at risk of post-operative death, and a related prediction tool or algorithm would help inform decisions for patients subjectively at risk of complications. For example, a high-risk patient may opt-out of surgery as the long-term benefits are outweighed by the cost. Providing a short-term mortality risk model could help inform decision making regarding whether to opt for the surgery and to help target preventative interventions.

In order to be clinically useful, covariates included in any model must be readily available at the time of model implementation. For this study this means pre-operatively. Current prediction model studies of post-operative outcomes after TKR have several limitations. In a recent review predicting post-operative infection after total joint replacement (4), most models were not externally validated, the process of applying a model in a new database to check if performance transfers to new data, and none were ready for clinical use due to issues with application (e.g. variables unobtainable at time of use) or insufficient performance. Some models were developed using data that were not routinely collected in observational data (e.g., floor of a patient's bedroom, preoperative walking distance) and therefore validation of these models was infeasible using the data available in this study. Finally, most models had not taken full advantage of all data available in medical records. For example, using a comorbidity index (5) instead of all patient characteristics (6). There is currently no TKR specific mortality prediction model.

A well performing robust model that predicts mortality could be used to aid in decision making for TKR as well as targeting interventions for high risk patients. As such the hypothesis of this study is that 90-day all-cause mortality is predictable using routinely collected data. This will be assessed by developing and externally validating a model using area under receiver operator curve.

MATERIALS AND METHODS

This retrospective cohort study used observational healthcare databases from the UK (The Health Improvement Network (THIN) (7)) and US (Optum). Detailed information on these databases is available in Table . All databases used in this paper were mapped into the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) (8) . The OMOP-CDM was developed for researchers to transform diverse datasets into a consistent structure and vocabulary. This means studies using these databases are more replicable increasing the clinical relevance of evidence.

Table 1 Data sources formatted to the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) used in this research (data type: claims, electronic health/medical records (EHR/EMR), general practitioner (GP))

Database	Database Acronym	Country	Data type	Time period
Optum© De-Identified Clinformatics® Data Mart Database	ClinFormatics	US	Claims	2000-2018
IQVIA Medical Research Data([IMRD], incorporating data from The Health Improvement Network [THIN]	THIN	UK	General Practice	2003-2018

Each site obtained institutional review board approval for the study or used de-identified data and therefore the study was determined not to be human subjects research. Informed consent was not necessary at any site.

Cohorts

Development Target Population Cohort

The target population for model development and validation contained patients with knee osteoarthritis undergoing TKR. The first recorded TKR procedure identified was considered the event of interest with the date of surgery as index date. Inclusion criteria required patients to have at least 1 year of continuous pre-index date recorded observation time. Individuals below the age of 40, those with prior evidence of knee arthroplasty, knee fracture, knee surgery (except diagnostic procedures), rheumatoid arthritis, inflammatory arthropathies, or septic arthritis at any time before the index date. This is because these patients likely have a cause other than osteoarthritis for their surgery. Patients with spine, hip, or foot pathology observed in the 365 days before index date were also excluded.

The target cohort for TKR is available at: TKR: http://atlas-demo.ohdsi.org/#/cohortdefinition/1776551

Outcome Cohorts

Mortality was defined as all-cause mortality based on records of date of death. This is well captured in THIN and in Optum until 2013, when a change in reporting means that the capture after this time is specific but less sensitive. Available at: http://atlas-demo.ohdsi.org/#/cohortdefinition/1776555

Patients were considered at risk for mortality from the day after surgery up until day 90.

Candidate Predictors

89,031 candidate predictors were derived from the observational healthcare data that existed on or prior to the target index date (TKR surgery date). These variables were demographics, binary indicators of medical events (e.g. GP visit, disease diagnosis, medication prescription) and counts of record types. The demographics were gender, 5 year age groups (40-45, 45-50,…,95+) and month of the target index date. Binary indicator variables for medical events were created based on the presence or absence of each concept for a patient corresponding to the OMOP-CDM clinical domains of conditions, drugs, procedures or measurements. For conditions binary predictors were created using the 30 days and 365 days prior to index date. For example, there exists one covariate for each of 'Diabetes mellitus', 'Hypertensive disorder', and 'Hypercholesterolemia' (and similarly for other diseases that appear in the patient records), based on the occurrence of a diagnosis code for each condition in the 365 days or 30 days preceding the index date. Drug covariates were constructed similarly, but used time windows of 30, 365, 1095 days and all time prior to target index date. Covariates representing counts how many visits (e.g. primary care visit) a patient had in the 365 days and 1095 days prior to the target index date were also created. The following existing risk scores (CHADS2, CHA2DS2VASc (both stroke risk models), Diabetes Complications severity index, Charlson Comorbidity Index) using all data prior to index were also calculated and used as candidate predictors.

Methodology for model development and validation

The study was initially conducted using the THIN and OPTUM datasets. Models predicting the 90-day mortality in the TKR target population were developed in both databases. The interoperability of the OMOP-CDM was utilised to externally validate in the non-development database.

Model development followed the framework for the creation and validation of patient-level prediction (PLP) models presented in Reps et al (9) , a person 'train-test split' method was used to perform internal validation. In each development cohort, the random split sample (`training sample') containing 75% of patients was used to develop the prediction models and the remaining 25% of patients (`test sample') was used to validate the risk scores. The models were trained using least absolute shrinkage and selection operator (LASSO) regularised logistic regression, using a 3-fold cross validation technique in the training sample to learn the optimal regularisation hyper-parameter through an adaptive search (10). LASSO regularization (11) helps to limit overfitting in model development. This process works by assigning a "penalty" to the inclusion of a variable, this variable must then contribute more to the performance than the penalisation. If this condition is not met then the coefficient of the covariate becomes 0, which eliminates the covariate from the model, thus automating feature selection.

Performance of the model was assessed in terms of discrimination and calibration. Discrimination assesses how well the model can distinguish which patients experience the outcome and calibration assesses whether the predicted risks are in alignment with the observed risks. Discrimination was measured using the Area Under Receiver Operator Characteristic Curve (AUROC). An AUROC of greater than 0.70 is considered to be a reasonable candidate for external validation. The model calibration was assessed by plotting the predicted and observed risks across deciles of predicted risk. Calibration assessment is then performed visually rather than using a statistic or numeric value as this provides an impression of the direction and scale of miscalibration (12) . Summary statistics were reported from the test samples.

External validation (13) was performed by applying the final prediction models in the dataset not used for development. The external validation was analysed in the same way as internally.

Model Parsimonisation

When using a data-driven approach to model development, generally the final models contain a large number of covariates. The full model assesses what is in principle the best possible performing model. However, the large number of covariates can create a barrier to implementation and understanding.

Models were therefore created that could be candidates for the clinical implementation by performing further analyses in order to reduce the number of features in the final model (improving parsimony). This analysis investigated what the performance loss is when using fewer covariates.

The approach involved analyzing the covariates selected by the final model and then using clinical expertise to attempt to combine multiple of these covariates, that correspond to a similar illness, into a single covariate. Often, LASSO logistic regression models include multiple covariates which are clinically related, for example a model might select the same condition occurrence but in different time periods predating the index date (e.g., 'diabetes -30 days to 0 days prior to index' and 'diabetes -365 days to 0 days prior to index'). These could be simplified to an aggregate covariate of "History of Diabetes", rather than multiple covariates specifying the specific time frame of the occurrence.

The procedures for developing both the full and parsimonious models will be identical except for the covariates. Definitions of the aggregated covariates are available in online Appendix 2.

All statistical analysis was performed using R (version 3.5.1) and the Patient-Level Prediction. This study was conducted and reported according to the Transparent Reporting of a multivariate prediction model for Individual Prediction or Diagnosis (TRIPOD) guidelines (14) . All the analysis code used for the development for the models are available on github at https://github.com/OHDSI/StudyProtocolSandbox/tree/master/mortalityValidation

as well as the developed mortality models themselves for external validation at: https://github.com/ohdsi-studies/TkrPredictSimple

RESULTS

The target population included 40,950 (THIN) and 152,665 (Optum) patients. 90-day mortality occurred in 0.20% (THIN)-0.23% (Optum) of patients (Table 2).

Table 2 TKR target and outcome population sizes and the internal AUROC achieved

Dataset	Target Population	90-day mortality	
		Size	AUROC
OPTUM	152,665	353 (0.23%)	0.78
THIN	40,950	81 (0.20%)	0.68

The 90-day mortality model trained using OPTUM obtained internal AUROC above 0.7 (Table 3). The external validation of the 90-day mortality models developed on OPTUM and THIN ranged between 0.68 to 0.86 and are presented in Table 4. Details of the distribution of key covariates can be found in online Appendix 1.

The OPTUM 90-day mortality model performed better than the THIN 90-day mortality model both internally and across the external validation (Table 2). The OPTUM 90-day mortality model achieved a slightly increased performance (AUROC 0.69) in the THIN dataset compared to the internal validation of the THIN developed model (AUROC 0.68). For the 90-day mortality OPTUM model, 102 of 89,031 candidate variables were selected into the final model. The full model is available in online Appendix 3.

The models and performance on the test and external validation sets are available to explore interactively at http://data.ohdsi.org/TKROutcomesExplorer/

The prevalence of a selection of covariates included in the 90-day mortality model developed using OPTUM, when assessed in multiple databases can be found in online Appendix 1.

This analysis shows that the covariate prevalence varies between the different databases, suggesting the databases have different underlying characteristics. As the models maintain performance despite these differences, it suggests that the model is robust to variability in the distribution of the covariates.

The 90-day Optum mortality was then parsimonised. The creation of these aggregate covariates and their definitions are available in online Appendix 2. This model is detailed in Table

When the analysis was performed with these covariates, the AUROC was 0.77 internally and 0.71 in THIN. The calibration plot for the internal validation and the THIN validation are presented in Figure 1. Figure 1 shows that, for the majority of patients, the model is well calibrated internally with the ideal line always appearing within the confidence interval. For the external validation in THIN, the model is well calibrated however for patients at higher risk there is some overestimation of risk in the highest risk groups. For example, a predicted risk of 0.02 corresponds to an observed risk of 0.015. The model could potentially benefit from recalibration in this setting.

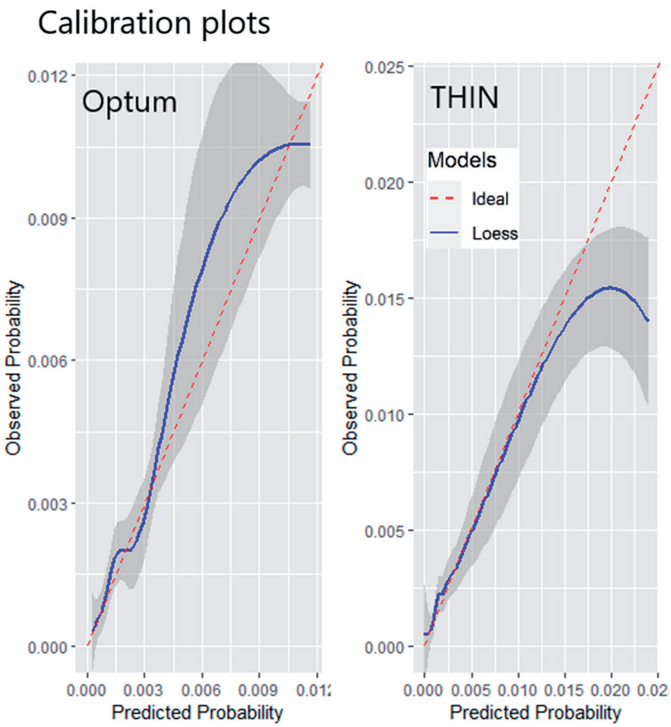


Figure 1 Calibration plot showing the calibration of the parsimonious model internally (Optum) and externally (THIN). The plot shows the agreement between the observed and predicted risk for patients. This is calculated by fitting loess regression.

DISCUSSION

The main finding of this study is the predictability of post-operative 90-day mortality following TKR. The AUROC of LASSO logistic regression model was found to be 0.78 in the OPTUM database. Validating this model against the other databases resulted in AUROC values of 0.68 (THIN) indicating that the model is fairly robust. The high number of features (102) in this model presents a barrier to implementation in clinics. A parsimonious model was therefore created, containing 12 variables. This model achieved AUROC of 0.77 in the training data and 0.71 in the external validation in the THIN database. The calibration was adequate although there appeared to be an overestimation of risk for patients at higher risk when assessed in THIN. As the parsimonious model achieved similar or better performance and is more implementable, it is preferred.

The desired operating characteristics when applying the parsimonious OPTUM 90-day mortality model to classify patients into those who will die and those who will not within 90 days of the surgery can be picked based on the prediction threshold, see Table 3. As an example,

Table 3 The covariates and values for the parsimonious 90-day mortality model

Covariate	Value
Intercept	-6.64376
Age Group	
40-44	-4.40718
45-49	-5.72523
50-54	-0.61149
55-59	-0.25853
60-64	-0.21392
65-69	-0.01862
70-74 (reference)	0
75-79	0.60808
80-84	1.08846
85-89	1.88595
90-94	-1.42352
Gender	
Male	0.36173
Female (reference)	0
History of:	
Cancer (excl non-melanoma skin cancer)	-0.21177
COPD	0.44467
Gout	0.45821
Heart Failure or Atrial Fibrillation	1.25532
Hypertension	-0.12567
Kidney disease	0.5571
OA	-0.4513
T2DM	0.27827
Opioid use	-0.35781
Psycholeptics use	0.17227

Table 4 The external validation of the 90-day mortality models

Development database	Validation database	Model Type	AUROC	Test population	Outcome count in test population (incidence in cases per 100 patients)
OPTUM	OPTUM	Full	0.78	38,166	88 (0.23)
OPTUM	THIN	Full	0.7	57,897	121 (0.30)
THIN	THIN	Full	0.68	10,237	20 (0.20)
OPTUM	OPTUM	Reduced	0.77	38,157	88 (0.23)
OPTUM	THIN	Reduced	0.71	57,897	121 (0.30)
THIN	OPTUM	Full	0.68	152,665	353 (0.23)

if a female patients aged 75 presented to a clinician whilst she had COPD and T2DM, then her raw score would be

–6.64376 (intercept) + 0.60808 (age = 75) + 0.44467 (COPD) + 0.27827 (T2DM) = –5.31274

which maps to a predicted risk of 0.5%. When compared to the outcome prevalence of 0.2% this shows the patient is twice as likely as average to die following this surgery.

In contrast to previous studies, the focus of this research was to develop the best performing predictive model on basis of all clinical and demographic data recorded in the observational databases and to then assess how close to this performance a reduced feature set model could come. The predictors included in the final model were mostly already known to be related to the outcome, what this study adds is to provide a quantitative relationship between the combination of these and the probability of the outcome. This was done by performing a regression analysis using these covariates. The selection of these predictors speaks to the robustness of the methods. Previous prediction models in the context of knee replacement have focused on patient-reported outcomes or revision surgery/implant survivorship, with little focus on complications or post-operative mortality, meaning comparison to these is difficult (15). When considering common mortality predictors such as the American College of Surgeons National Surgical Quality Improvement Program comparisons are difficult using observational data as "Functional status" are not well captured in observational studies. Further, the Revised Cardiac Risk Index generally performs with a median AUROC of 0.62 showing lower performance than the model developed in this study (16).

Hunt et al. report an incidence of mortality (0.37%) in their study on 45-day mortality following knee replacement surgery (17). This is high compared with our reported incidence of mortality, which could be due to the limitation of the mortality capture in the databases studied. The low incidence of death (around 0.2%) following TKR necessitates large datasets with accurate recording of mortality. The reported 90-day mortality predictive model may be used as a complementary element for screening of high-risk patients and better preparation before surgery. It could also allow the patient and clinician to be better informed about the potential benefit-risk of elective TKR. Given that all-cause mortality was considered, the mortality is not necessarily caused by the TKR, however if the patient is deemed to be at a high risk of mortality in the 90-day post-operative period then the surgery is still likely inadvisable due to the costs to both the patient and the healthcare system without providing benefit.

Limitations of this study include the low number of outcomes in some of the analyses meaning that estimates are potentially unreliable, as well as potential misclassification of covariates in the data. The recording of death in the THIN is very reliable but in Optum is known to be specific but lacking some sensitivity because in 2013 reporting of death stopped being mandatory. This could lead to an underestimation of the number of deaths following a TKR in this study. Further limitations are that although large numbers of covariates are included in the analysis, some covariates are poorly captured in the data used. Known predictors such as surgeon skill and volume are not available in routinely collected healthcare data and as such have not been included. As with all observational studies, the models can only be assessed on the predictors available and as such any predictors which are not in the source data, will be missed by the models.

Limitations of the phenotypes include: i) there is a potential contamination issue in the TKR cohort as prior to ICD-10 coding, TKR cohorts will have UKR cases as the same ICD procedure code was valid for both ii) if a patient were to have bilateral TKR only the first surgery would be included in our target cohort and the second would be excluded.

A major strength of this study is that the model is already externally validated, demonstrating its robustness and transportability, a process typically taking 3-years (12). The low number of features of this model is a significant advantage to implementation.

CONCLUSION

In conclusion, models were developed and externally validated for 90-day mortality after a TKR prediction model that has both good discrimination performance and calibration which are maintained across the external validation. Thus, this model is a strong candidate for impacting clinical decision making.

Supplementary Information

The online version contains supplementary material available at https://rdcu.be/digIZ

Acknowledgements

The authors would like to thank the OHDSI community for their contributions to the tools used for this analysis. The authors would also like to thank Paloma O'Dogherty Cordero who helped to organise the Oxford OHDSI study-a-thon, from which this study emerged. IQVIA Medical Research Data incorporates data from The Health Improvement Network, THIN. THIN is a registered trademark of Cegedim SA in the UK and other countries. Reference made to the THIN database is intended to be descriptive of the data asset licensed by IQVIA. This work uses de-identified data provided by patients as a part of their routine primary care.

Thanks to Anthony G. Sena BA, Edward Burn MSc, Ying He PhD, Daniel R. Morales PhD, David Culliford PhD, Dahai Yu PhD’ Victoria Y. Strauss PhD, Talita Duarte-Salles PhD, Albert Prats-Uribe MPH, Antonella Delmestri PhD, James Weaver MSc’ William Sproviero PhD, Danielle Robinson MSc, Henry Morgan Stewart PhD’ Belay Birlie MSc, Rafael Pinedo-Villanueva PhD, Spyros Kolovos PhD, Luis H. John MSc, Ruth E. Costello MSc, Michel van Speybroeck MSc, Caroline O’Leary MSc, Evan Minty MD MSc, Thomas Falconer MS, Alison Callahan PhD, Stephen Pfohl BS, Theresa Burkard PhD, Jennifer Lane MEd MRCS all of whom contributed to The OHDSI/EHDEN Knee Arthroplasty Group, without whom this work would have been impossible.

Funding

This activity under the European Health Data & Evidence Network (EHDEN) has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA. The sponsor of the study did not have any involvement in the writing of the manuscript or the decision to submit it for publication. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Competing interests

All authors have completed the ICMJE uniform disclosure form at www.icmje.org/coi_disclosure.pdf and declare:

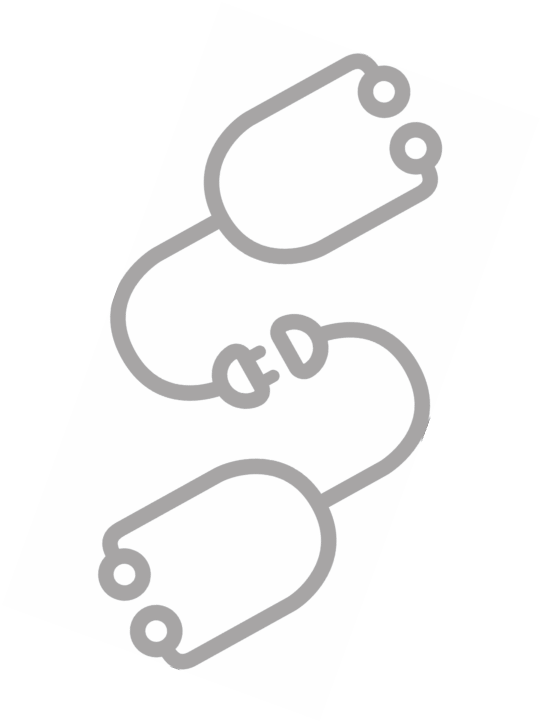
AS, JW, JR, MvS and PBR are full-time employees of Janssen Research & Development, a pharmaceutical company of Johnson & Johnson, and shareholders in Johnson & Johnson. The Johnson & Johnson family of companies also includes DePuy Synthes, which is the maker of medical devices for joint reconstruction. DPA reports grants from Amgen, grants from UCB Biopharma, grants from Les Laboratoires Servier, outside the submitted work. CO is a part-time employee of IQVIA.

Author contributions

All authors made substantial contributions to the conception or design of the work; RW, PR, DPA and JL constructed the aggregate covariates DPA and PBR led the acquisition of the data; all authors were involved in the analysis and interpretation of data for the work; All authors have contributed to the drafting and revising critically the manuscript for important intellectual content; all authors have given final approval and agree to be accountable for all aspects of the work.

BIBLIOGRAPHY

1. Arden N, Altman D, Beard D, Carr A, Clarke N, Collins G, et al. *Lower limb arthroplasty: can we produce a tool to predict outcome and failure, and is it cost-effective? An epidemiological study*;10.3310/pgfar05120. Southampton (UK)2017.
2. Berstock JR, Beswick AD, Lopez-Lopez JA, Whitehouse MR, Blom AW (2018) Mortality After Total Knee Arthroplasty: A Systematic Review of Incidence, Temporal Trends, and Risk Factors. J Bone Joint Surg Am 100:1064-1070
3. Blak BT, Thompson M, Dattani H, Bourke A (2011) Generalisability of The Health Improvement Network (THIN) database: demographics, chronic disease prevalence and mortality rates. Inform Prim Care 19:251-255
4. Ford MK, Beattie WS, Wijeysundera DN (2010) Systematic review: prediction of perioperative cardiac complications and mortality by the revised cardiac risk index. Ann Intern Med 152:26-35
5. Hunt LP, Ben-Shlomo Y, Clark EM, Dieppe P, Judge A, MacGregor AJ, et al. (2014) 45-day mortality after 467,779 knee replacements for osteoarthritis from the National Joint Registry for England and Wales: an observational study. Lancet 384:1429-1436
6. Inacio MCS, Pratt NL, Roughead EE, Graves SE (2016) Evaluation of three co-morbidity measures to predict mortality in patients undergoing total joint arthroplasty. Osteoarthritis Cartilage 24:1718-1726
7. Iqbal J, Vergouwe Y, Bourantas CV, van Klaveren D, Zhang YJ, Campos CM, et al. (2014) Predicting 3-year mortality after percutaneous coronary intervention: updated logistic clinical SYNTAX score based on patient-level data from 7 contemporary stent trials. JACC Cardiovasc Interv 7:464-470
8. Konopka JF, Hansen VJ, Rubash HE, Freiberg AA (2015) Risk assessment tools used to predict outcomes of total hip and total knee arthroplasty. Orthop Clin North Am 46:351-362, ix-x
9. Kunutsor SK, Whitehouse MR, Blom AW, Beswick AD (2017) Systematic review of risk prediction scores for surgical site infection or periprosthetic joint infection following joint arthroplasty. Epidemiol Infect 145:1738-1749
10. Moons KG, Altman DG, Reitsma JB, Ioannidis JP, Macaskill P, Steyerberg EW, et al. (2015) Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med 162:W1-73
11. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE (2012) Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc 19:54-60
12. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, et al. (2012) Mortality after surgery in Europe: a 7 day cohort study. Lancet 380:1059-1065
13. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR (2018) Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc;10.1093/jamia/ocy032
14. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. (2020) Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol 20:102
15. Springer BD, Cahue S, Etkin CD, Lewallen DG, McGrory BJ (2017) Infection burden in total hip and knee arthroplasties: an international registry-based perspective. Arthroplast Today 3:137-140
16. Suchard MA, Simpson SE, Zorych I, Ryan P, Madigan D (2013) Massive parallelization of serial inference algorithms for a complex generalized linear model. ACM Trans Model Comput Simul 23:10
17. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Series B Stat Methodol 58:267-288



7

Development and external validation of prediction models for adverse health outcomes in rheumatoid arthritis: a multinational real-world cohort analysis

Cynthia Yang[1*#], MSc; Ross D. Williams[1*], MSc; Joel N. Swerdel[2], PhD; João Rafael Almeida[3], PhD; Emily S. Brouwer[2], PhD; Edward Burn[4,5], PhD; Loreto Carmona[6], MD, hD; Katerina Chatzidionysiou[7], MD, PhD; Talita Duarte-Salles[5], PhD; Walid Fakhouri[8], PhD; Antje Hottgenroth[9], PhD; Meghna Jani[10], MRCP, PhD; Raivo Kolde[11], PhD; Jan A. Kors[1], PhD; Lembe Kullamaa[12,13,14], MSc; Jennifer Lane[4], MRCS; Karine Marinier[15], MSc; Alexander Michel[16], MD, PhD; Henry Morgan Stewart[17], PhD; Albert Prats-Uribe[4], MD; Sulev Reisberg[11,18,19], PhD; Anthony G. Sena[1,2], BA; Carmen O. Torre[17], MSc; Katia Verhamme[1], MD, PhD; David Vizcaya[20], PhD; James Weaver[2,21], MSc; Patrick Ryan[2,21], PhD; Daniel Prieto-Alhambra[4], MD, PhD; Peter R. Rijnbeek[1], PhD

* Cynthia Yang and Ross D. Williams contributed equally to this paper.
[#] Corresponding author Cynthia Yang, Dr. Molewaterplein 40, 3015 GD Rotterdam, c.yang@erasmusmc.nl, https://orcid.org/0000-0001-6769-3153.

[1]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands; [2]Janssen Research and Development, Titusville, NJ, USA; [3]DETI/IEETA, University of Aveiro, Aveiro, Portugal; [4]Nuffield Department of Orthopaedics Rheumatology and Musculoskeletal Sciences, University of Oxford, Oxford, UK; [5]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain; [6]Instituto de Salud Musculoesquelética, Madrid, Spain; [7]Dep of Medicine, Solna, Rheum Unit, Karolinska Institute, Stockholm, Sweden; [8]Eli Lilly and Company, Windlesham, Surrey, UK; [9]Lilly Deutschland GmbH, Bad Homburg, Germany; [10]Centre for Epidemiology Versus Arthritis, University of Manchester, Manchester, UK; [11]Institute of Computer Science, University of Tartu, Tartu, Estonia; [12]Department of Epidemiology and Biostatistics, National Institute for Health Development, Tallinn, Estonia; [13]Institute of Family Medicine and Public Health, University of Tartu, Tartu, Estonia; [14]European Patients' Forum, Brussels, Belgium; [15]Servier, Suresnes, France; [16]Epidemiology, Bayer Basel, Basel, Switzerland; [17]Real-World Solutions, IQVIA, Brighton, UK; [18]STACC, Tartu, Estonia; [19]Quretec, Tartu, Estonia; [20]Bayer Pharmaceuticals, Barcelona, Spain; [21]Observational Health Data Sciences and Informatics, New York, NY, USA

Link to article: https://pubmed.ncbi.nlm.nih.gov/35728447/

ABSTRACT

Background: Identification of rheumatoid arthritis (RA) patients at high risk of adverse health outcomes remains a major challenge. We aimed to develop and validate prediction models for a variety of adverse health outcomes in RA patients initiating first-line methotrexate (MTX) monotherapy.

Methods: Data from 15 claims and electronic health record databases across 9 countries were used. Models were developed and internally validated on Optum® De-identified Clinformatics® Data Mart Database using L1-regularized logistic regression to estimate the risk of adverse health outcomes within 3 months (leukopenia, pancytopenia, infection), 2 years (myocardial infarction (MI) and stroke), and 5 years (cancers [colorectal, breast, uterine]) after treatment initiation. Candidate predictors included demographic variables and past medical history. Models were externally validated on all other databases. Performance was assessed using the area under the receiver operator characteristic curve (AUC) and calibration plots.

Findings: Models were developed and internally validated on 21,547 RA patients and externally validated on 131,928 RA patients. Models for serious infection (AUC: internal 0.74, external ranging from 0.62 to 0.83), MI (AUC: internal 0.76, external ranging from 0.56 to 0.82), and stroke (AUC: internal 0.77, external ranging from 0.63 to 0.95), showed good discrimination and adequate calibration. Models for the other outcomes showed modest internal discrimination (AUC < 0.65) and were not externally validated.

Interpretation: We developed and validated prediction models for a variety of adverse health outcomes in RA patients initiating first-line MTX monotherapy. Final models for serious infection, MI, and stroke demonstrated good performance across multiple databases and can be studied for clinical use.

Funding: This activity under the European Health Data & Evidence Network (EHDEN) has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA.

INTRODUCTION

Compared to the general population, patients with rheumatoid arthritis (RA) have an increased risk of treatment-related adverse events, such as cytopenia and infection, and comorbidities, such as cardiovascular disease (CVD) and cancer (1-3). Although the management and prognosis of RA has improved in recent decades, identification of RA patients at high risk of adverse health outcomes remains a major challenge (4, 5).

The European Alliance of Associations for Rheumatology (EULAR) and the American College of Rheumatology (ACR) recommend initiating methotrexate (MTX) monotherapy (with glucocorticoids) as soon as possible after the diagnosis of RA (6, 7), making this the most common treatment for RA worldwide. MTX treatment implies screening or monitoring efficacy and side-effects, as with most disease modifying antirheumatic drugs (DMARDs). Using prediction models to evaluate patient-level risks in RA patients initiating first-line MTX monotherapy could allow clinicians to target those at high risk of adverse health outcomes for increased screening or monitoring throughout the course of treatment.

Few prediction models have been developed for adverse health outcomes in RA patients, with those that have been developed focusing on the risk of either CVDs or serious infection (8-15). Challenges in the development of RA-specific prediction models have previously been highlighted (10, 16). For example, while existing CVD models estimate 10-year risks, a shorter period may be more appropriate, since most RA patients will change treatments several times during a 10-year period (10). Additionally, a larger cohort of RA patients would allow for the development of RA-specific prediction models using a larger number of candidate predictors (8). Finally, most existing models have not been subjected to extensive external validation, which is necessary to understand a model's prediction performance (17). In this study, we aimed to develop and externally validate prediction models for a variety of adverse health outcomes in RA patients initiating first-line MTX monotherapy, using 15 large-scale claims and electronic health record (EHR) databases across 9 countries and 4 continents.

MATERIALS AND METHODS

This study was conducted within the European Health Data & Evidence Network (EHDEN) project and involved a multidisciplinary team of rheumatologists, clinicians, epidemiologists, data custodians, and data scientists. We developed and validated prediction models using the Patient-Level Prediction framework from the Observational Health Data Sciences and Informatics (OHDSI) initiative (18). This framework allows for standardized development and extensive validation of prediction models using observational health databases and can be applied to datasets that are mapped to the Observational Medical Outcomes Partnership Common Data Model (OMOP CDM) (19-21). The OMOP CDM was developed to transform source data into a common format and enables analytical source code to be shared among researchers. We followed the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) guidelines for reporting (22).

Data sources

We used 15 claims and EHR databases with data mapped to the OMOP CDM from 6 European countries (Spain, Estonia, Netherlands, Germany, France, and the United Kingdom (UK)), the United States of America (USA), Australia, and Japan. The databases are listed in Table 1. Data from the Optum® De-Identified Clinformatics® Data Mart Database, a USA claims database, were used for model development and internal validation. Data from the 14 other databases were used for external validation. Each site obtained institutional review board approval for the study or used de-identified data. Therefore, informed consent was not necessary at any site. Extended descriptions of the databases are provided in Table A.1 in online Appendix A.

Study population

Adult RA patients (aged 18 years and over) were included in the study population if they had at least 365 days of observation in the database prior to the first drug utilization record of MTX (the index event) and met all of the following inclusion criteria: 1) a diagnosis of RA within 5 years prior to or on index, 2) no drug utilization record of any DMARD any time prior to index, 3) no drug utilization record of any other DMARD on or within 7 days after index, 4) no record indicating any cancer any time prior to or on index, and 5) no record indicating any other inflammatory arthritis (psoriatic arthritis, ankylosing spondylitis, reactive arthritis, any axial spondyloarthropathy) any time prior to or on index.

Detailed definitions of these inclusion criteria, including code lists, are available at http://atlas-demo.ohdsi.org/#/cohortdefinition/1773112.

Outcomes

We investigated outcomes for which RA patients have increased risks compared to the general population and for which RA patients identified at high risk could be targeted for increased screening or monitoring throughout the course of treatment. The first event (binary) of each of the following adverse health outcomes within a period after initiating first-line MTX monotherapy (the index event) was considered: 1) leukopenia, pancytopenia, and infection (serious, opportunistic, all) recorded from 1 day up to 90 days after index, 2) myocardial infarction (MI) and stroke recorded from 1 day up to 2 years after index, 3) cancer (colorectal, breast, uterine) recorded from 1 year up to 5 years after index.

Detailed definitions of these outcomes, including code lists, are available at: https://github.com/ohdsi-studies/EhdenRaPrediction/tree/master/inst/cohorts.

For all outcomes, patients who had any record of the specific outcome within 90 days prior to or at initiation of MTX monotherapy were excluded from the study population. For

Table 1. Databases included in the study with data mapped to the OMOP CDM

Database full name	Database short name	Country	Data type	Patient type	Population size	Data range
Estonian Health Information System	Estonia	Estonia	EHR	All inpatient and outpatient discharge summaries, general population	1.4m	2012-2016
IBM MarketScan® Commercial Database	CCAE	USA	Claims	Privately insured	151m	2000-2020
IBM MarketScan® Medicare Supplemental Database	MDCR	USA	Claims	Retiree supplemental	10m	2000-2020
IBM MarketScan® Multi-State Medicaid Database	MDCD	USA	Claims	Medicaid	30m	2006-2020
Integrated Primary Care Information	IPCI	Netherlands	EHR	Primary care	2.6m	1996-2020
IQVIA Australia EMR	IQVIA Australia	Australia	EHR	Outpatient / General population	6m	2006-2020
IQVIA Disease Analyser Germany EMR	IQVIA Germany	Germany	EHR	Outpatient / General population Public and private insurance	37m	1992-2020
IQVIA Hospital US Charge Master	IQVIA US Hospital	USA	EHR	Inpatient & outpatient hospital encounters, including Emergency Room visits / General population	86m	2007-2020
IQVIA LPD France	IQVIA LPD France	France	EHR	Outpatient / General population	7.8m	1994-2020
IQVIA UK THIN IMRD EMR	IQVIA THIN	UK	EHR	General population / Primary care records with hospitalization / referral information	15m	1989-2020
IQVIA US Ambulatory EMR	IQVIA US Ambulatory	USA	EHR	Outpatient / General population	49m	2006-2020
Japan Medical Data Center	JMDC	Japan	Claims	Society-Managed Health Insurance	12m	2005-2020
Optum® De-Identified Clinformatics® Data Mart Database	Optum Claims	USA	Claims	Privately insured	87m	2001-2020
Optum® De-identified Electronic Health Record Dataset	Optum EHR	USA	EHR	Privately insured	100m	2006-2020
The Information System for Research in Primary Care	SIDIAP-H	Spain	EHR	Primary care linked (partially) to inpatient data	5.8m	2006-2020

cancer outcomes, patients who were lost to follow-up within one year of treatment initiation were excluded. For all other outcomes, patients who were lost to follow-up within one day of treatment initiation were excluded. Outcomes for which the final study population contained less than 25 RA patients with an outcome event were omitted from further analysis.

Candidate predictors

Candidate predictors were extracted from data routinely recorded in the database. This included binary indicators of 5-year age groups (i.e., 20-24, 25-29, etc.) and sex, as well as a large set of binary indicators of recorded OMOP CDM concepts for health conditions and drug groups (44). For health conditions, we considered all data prior to index. For drug groups, we separately considered data from the 30 days prior to index and data from the 365 days prior to index. Finally, three established risk scores (CHA2DS2-VASc, Diabetes Complications Severity Index (DCSI), Charlson Comorbidity Index (CCI) (Romano adaptation)) were calculated using all data prior to index (23-26). No clinical measurements or sporadically recorded variables were included as candidate predictors to maximize transportability (i.e., ability to apply across databases) of the developed prediction models.

Handling of missing data

In the observational data used in this study, if a candidate predictor was not recorded in a patient's history, we assumed that the candidate predictor was not observed for this patient. Age group and sex are required by the OMOP CDM and were always recorded. For our analyses, if a health condition or drug group was not recorded in a patient's history, we assumed that the health condition or drug group was absent.

Statistical analysis methods

We used logistic regression with predictor selection through L1-regularization (27). For each outcome, two L1-regularized logistic regression models were developed: 1) one model using all candidate predictors for a data-driven approach of predictor selection, and 2) one model using only age groups and sex as candidate predictors to provide a benchmark.

A random subset of 75% of the patients was used as a training set and the remaining subset of 25% of the patients was used as a test set. First, 3-fold cross-validation was performed on the training set to optimize the regularization parameter, after which the test set was used for internal validation. Discrimination was assessed numerically using the area under the receiver operator characteristic curve (AUC) with a 95% confidence interval (CI). Calibration was assessed graphically by plotting the predicted risks against the observed risks.

To examine model performance across multiple databases, we externally validated the models. Only models with an AUC of at least 0.7 on internal validation were considered good enough to warrant external validation. External validation of each model was limited to those databases within which the corresponding outcome events could be identified. Discrimination and calibration were assessed in the same way as on internal validation.

Software packages containing the analytical source code that was used to develop the models and to externally validate the developed models on databases with data mapped to the OMOP CDM are available at https://github.com/ohdsi-studies/EhdenRaPrediction.

RESULTS

Study population

In the development database (Optum Claims), 21,547 RA patients met the inclusion criteria. For breast cancer and uterine cancer, we only included female patients, resulting in 15,311 RA patients who met the inclusion criteria. Table 2 shows the number of RA patients and the number of RA patients with an outcome event in the final study population for each adverse health outcome. An attrition flowchart explaining how we arrived at the number of patients in the final study population for each outcome is available in Figure B.1 in online Appendix B.

Table C.1. in online Appendix C shows demographics and baseline characteristics of the final study population for each outcome, based on all data prior to or at initiation of MTX monotherapy. Patients with an outcome event on average had more comorbidities at treatment initiation.

Model specification

A total of 12,724 candidate predictors were extracted from data routinely recorded in the database, of which 18 were binary indicators of 5-year age groups and sex.

The number of predictors in each final model is specified in Table 3. Full lists of candidate predictors and detailed specifications of all final models are available in an interactive R Shiny

Table 2. Final study population in the Optum Claims database

Outcome	Number of RA patients	Number of RA patients with outcome event (%)
Leukopenia	21,452	85 (0.4)
Pancytopenia	21,496	30 (0.1)
Serious infection	21,276	316 (1.5)
Opportunistic infection	21,404	161 (0.8)
All infection	19,163	1,957 (10.2)
Myocardial infarction	21,463	417 (1.9)
Stroke	21,425	527 (2.5)
Colorectal cancer	15,584	53 (0.3)
Breast cancer	11,072	100 (0.9)
Uterine cancer	11,104	18 (0.2)

Table 3. Internal discrimination of the models developed on the Optum Claims database

Outcome	Data-driven approach (age groups, sex, conditions, drugs)		Benchmark (age groups, sex)	
	Number of predictors	AUC (95% CI)	Number of predictors	AUC (95% CI)
Stroke	90	0.77 (0.73-0.81)	16	0.74 (0.70-0.78)
Myocardial infarction	64	0.76 (0.72-0.81)	16	0.72 (0.68-0.76)
Serious infection	87	0.74 (0.68-0.80)	13	0.68 (0.62-0.74)
Opportunistic infection	12	0.60 (0.51-0.68)	1	0.49 (0.42-0.57)
All infection	62	0.59 (0.57-0.62)	6	0.53 (0.50-0.56)
Colorectal cancer	1	0.55 (0.41-0.69)	2	0.64 (0.48-0.79)
Leukopenia	10	0.50 (0.36-0.64)	NA	NA
Breast cancer	NA	NA	4	0.52 (0.42-0.61)

web application at https://data.ohdsi.org/EhdenRaPrediction/. The candidate predictors can be explored interactively in the 'Model Table' under 'Model'; an overview of the predictors in the final model can also be exported from this tab using 'Download Model'.

Model performance

Model discrimination on internal validation is presented in Table 3, ordered by highest AUC for the data-driven approach. For leukopenia, the AUC was below 0.6, indicating modest discrimination. For pancytopenia, no predictors were identified for both the data-driven approach and the benchmark, and we were therefore unable to develop any prediction model for this outcome. For opportunistic infection, and all infection, the AUCs were 0.6 or lower. For serious infection, MI, and stroke, the data-driven approach resulted in AUCs on internal validation of 0.74 (0.68-0.80), 0.76 (0.72-0.81), and 0.77 (0.73-0.81), respectively, indicating good discrimination. For colorectal cancer and breast cancer, the AUCs were below 0.65. Finally, for uterine cancer, the final study population contained less than 25 RA patients with an outcome event. Therefore, this outcome was omitted from further analysis.

The calibration plots for serious infection, MI, and stroke (Figure D.1-D.6 in online Appendix D) indicated adequate calibration besides good discrimination. We externally validated the models for these three outcomes across the 14 other databases. The AUCs are presented in Table 4, ordered by highest AUC for the data-driven approach. Overall, the models demonstrated good performance across multiple databases. Several 95% CIs were wide due to limited statistical power, making those results difficult to interpret. We therefore focused on the databases within which at least 100 RA patients with an outcome event were identified, which is a recommended minimum for external validation (28). In these databases, the data-driven approach consistently outperformed the benchmark, with AUCs ranging from 0.62 to 0.76 for serious infection, from 0.65 to 0.75 for MI, and from 0.63 to 0.79 for stroke. The corresponding calibration plots from the data-driven approach are presented in Figure E.1-E.16 in online

Table 4. External discrimination of the models for serious infection, myocardial infarction, and stroke

Outcome	Database	Number of RA patients	Number of RA patients with outcome event (%)	Data-driven approach (age groups, sex, conditions, drugs) - AUC (95% CI)	Benchmark (age groups, sex) - AUC (95% CI)
Serious infection	Estonia	1,441	8 (0.6)	0.83 (0.64-1.00)	0.76 (0.66-0.86)
	SIDIAP-H	2,051	5 (0.2)	0.78 (0.65-0.92)	0.76 (0.58-0.94)
	Optum EHR	29,980	308 (1.0)	0.76 (0.73-0.79)	0.63 (0.60-0.66)
	JMDC	4,871	15 (0.3)	0.72 (0.62-0.82)	0.63 (0.51-0.75)
	MDCR	6,662	154 (2.3)	0.67 (0.62-0.71)	0.59 (0.55-0.64)
	CCAE	29,303	223 (0.8)	0.65 (0.61-0.68)	0.54 (0.51-0.58)
	MDCD	3,793	123 (3.2)	0.63 (0.58-0.68)	0.50 (0.44-0.55)
	IQVIA US Hospital	3,871	776 (20.0)	0.62 (0.59-0.64)	0.56 (0.54-0.58)
Myocardial infarction	IPCI	1,458	24 (1.6)	0.82 (0.72-0.92)	0.78 (0.72-0.84)
	Optum EHR	30,568	466 (1.5)	0.75 (0.73-0.77)	0.71 (0.69-0.73)
	IQVIA LPD France	2,652	5 (0.2)	0.74 (0.60-0.87)	0.50 (0.26-0.75)
	SIDIAP-H	2,057	17 (0.8)	0.72 (0.61-0.84)	0.65 (0.54-0.77)
	IQVIA US Ambulatory	28,129	114 (0.4)	0.72 (0.67-0.77)	0.66 (0.62-0.70)
	MDCD	3,876	88 (2.3)	0.69 (0.63-0.75)	0.61 (0.55-0.67)
	CCAE	29,509	186 (0.6)	0.68 (0.64-0.73)	0.65 (0.61-0.68)
	MDCR	6,739	218 (3.2)	0.67 (0.64-0.71)	0.57 (0.53-0.61)
	IQVIA Germany	8,046	39 (0.5)	0.67 (0.57-0.77)	0.51 (0.41-0.61)
	IQVIA US Hospital	4,331	190 (4.4)	0.65 (0.61-0.69)	0.60 (0.56-0.63)
	Estonia	1,455	18 (1.2)	0.62 (0.48-0.77)	0.71 (0.61-0.81)
	JMDC	4,899	12 (0.2)	0.59 (0.44-0.75)	0.59 (0.40-0.77)
	IQVIA THIN	7,850	51 (0.6)	0.59 (0.50-0.67)	0.64 (0.58-0.70)
	IQVIA Australia	359	7 (1.9)	0.56 (0.29-0.84)	0.62 (0.45-0.80)

Table 4. External discrimination of the models for serious infection, myocardial infarction, and stroke *(continued)*

Outcome	Database	Number of RA patients	Number of RA patients with outcome event (%)	Data-driven approach (age groups, sex, conditions, drugs) - AUC (95% CI)	Benchmark (age groups, sex) - AUC (95% CI)
Stroke	IPCI	1,467	6 (0.4)	0.95 (0.90-0.99)	0.82 (0.65-0.98)
	Optum EHR	30,506	517 (1.7)	0.79 (0.77-0.81)	0.73 (0.71-0.76)
	IQVIA Germany	8,052	41 (0.5)	0.78 (0.70-0.86)	0.65 (0.57-0.72)
	MDCD	3,864	129 (3.3)	0.78 (0.74-0.82)	0.69 (0.64-0.73)
	Estonia	1,455	23 (1.6)	0.75 (0.67-0.83)	0.72 (0.64-0.81)
	JMDC	4,899	32 (0.7)	0.75 (0.65-0.84)	0.64 (0.55-0.74)
	SIDIAP-H	2,057	23 (1.1)	0.72 (0.63-0.81)	0.71 (0.62-0.79)
	IQVIA US Ambulatory	28,163	115 (0.4)	0.71 (0.67-0.76)	0.68 (0.64-0.72)
	CCAE	29,509	259 (0.9)	0.70 (0.67-0.74)	0.64 (0.61-0.67)
	MDCR	6,734	304 (4.5)	0.67 (0.64-0.70)	0.57 (0.54-0.61)
	IQVIA THIN	7,852	23 (0.3)	0.66 (0.56-0.76)	0.65 (0.55-0.74)
	IQVIA US Hospital	4,306	208 (4.8)	0.63 (0.59-0.67)	0.62 (0.59-0.66)

Appendix E. These plots showed adequate calibration for all three outcomes. As expected, the models tend to underestimate or overestimate risk in databases with a higher or lower incidence, respectively (29). For example, the model for serious infection underestimated the risk in the IQVIA US Hospital database, where the outcome incidence was more than tenfold higher than the outcome incidence in the development database. To account for this, the models can be recalibrated for use in these databases.

The final data-driven models for serious infection, MI, and stroke, including intercept, coefficients, and OMOP CDM concept IDs, can be found in Table F.1-F.3 in online Appendix F. Several age groups corresponding to older age were selected as predictors. Sex was selected as predictor in the model for stroke. The CHADS2-VASc score was selected as predictor in both the model for MI and the model for stroke. Furthermore, within each model, a large set of binary indicators of conditions and drugs was selected as predictors.

DISCUSSION

In this study, we developed and validated prediction models for a variety of adverse health outcomes in RA patients initiating first-line MTX monotherapy. For serious infection, MI, and stroke, the models demonstrated good performance across multiple databases. Internal validation of these models resulted in AUCs of greater than 0.70 and adequate calibration. External validation of these models resulted in good discrimination, where the data-driven approach of predictor selection (age groups, sex, conditions, drugs) consistently outperformed the benchmark (age groups, sex). This shows that conditions and drugs extracted from routinely recorded data have added value in identifying patients at high risk of serious infection, MI, and stroke. The models showed adequate calibration as well, although for some databases, the models may benefit from recalibration.

For uterine cancer, we were not able to develop models using our data. For this outcome, more data are required. For leukopenia, pancytopenia, opportunistic infection, all infection, colorectal cancer, and breast cancer, we did not externally validate the developed models since they did not discriminate well (AUC < 0.65) on internal validation.

We developed our models using large-scale claims and EHR databases that contain routinely recorded patient information. We chose for a data-driven approach and considered all conditions and drugs in a patient's history as candidate predictors. It is interesting to note that this large set of conditions and drugs in a patient's history also included comedication with nonsteroidal anti-inflammatory drugs and glucocorticoids, which may capture aspects of RA-specific variables such as disease activity indicators that were not explicitly considered as candidate predictors. It is important to note that our study was focused on prediction and not on evaluating individual predictor associations; it may be misleading to highlight individual predictors as having predictive value by themselves, since this can lead to causal interpretation

(30). It could be interesting to investigate whether explicitly considering certain RA-specific variables as candidate predictors would improve prediction performance. A potential limitation of our study is that we were not able to investigate this using our data.

A potential limitation of our study is that routinely recorded data can be misclassified. If a candidate predictor or an adverse health outcome was not recorded in a patient's history, we assumed but could not be certain that there had not been such an event. Variation in prediction performance across databases may reflect differences in how data are captured. Another potential limitation of our study is that our models were developed for RA patients initiating first-line MTX monotherapy and are therefore only intended for this target population. However, since MTX is the current anchor drug in RA, the developed models are applicable to a large group of RA patients initiating first-line treatment. Furthermore, although it is possible that the models may be applicable to RA patients initiating other treatments, this has not been investigated in our research and would require further validation of the models in those target populations. Finally, a potential limitation of our study is that we did not perform any sensitivity analyses on the inclusion criteria used to obtain our study population. For example, only patients on MTX with no drug utilization record of any other DMARD on or within 7 days after index were included. The 7 days after index offset was chosen to avoid any other DMARDs that occurred at the index date but were entered late in the record. Although less likely, it is possible that a second DMARD that occurred at the index date was entered in the record with a delay longer than 7 days after index.

Our study also has several strengths. To the best of our knowledge, this study is the largest cohort study to date on predicting a variety of adverse health outcomes in RA patients. By developing models using data mapped to the OMOP CDM, we were able to include 14 databases for external validation of the models. The scale of the data allowed us to investigate a variety of adverse health outcomes and validate the models across multiple international databases. Even though more data are still needed for some of the outcomes, our study shows the feasibility of this approach. Additionally, we have provided software packages that contain the analytical source code used to develop the models and to externally validate our developed models on databases with data mapped to the OMOP CDM.

The models developed and internally validated for serious infection, MI, and stroke using a large-scale USA claims database (Optum Claims) showed good performance across multiple claims and EHR databases. We do not believe we have sufficient evidence to recommend the use of the models in clinical practice, but research could be conducted to prove their value as implemented in administrative or EHR software to generate automatic reminders for individuals at high risk of these outcomes. The models are based entirely on routinely recorded data, minimizing the burden to the clinician. However, regulatory approvals would be required before this can be considered, which is beyond the scope of our current work. The models had particularly good external validation performance on the USA databases and appeared to perform well (with wider confidence intervals) on the international databases too. RA patients identified at high risk of serious infection, MI, or stroke could be targeted for increased screening or monitoring throughout the course of treatment, complementary to current screening or monitoring strategies. In this way, the models may enable clinicians to provide better personalized care to RA patients initiating first-line MTX monotherapy.

Supplementary Information

The online version contains supplementary material available at https://pubmed.ncbi.nlm.nih.gov/35728447/

Funding

This activity under the European Health Data & Evidence Network (EHDEN) has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968. This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. The sponsor of the study did not have any involvement in the writing of the manuscript or the decision to submit it for publication. The research was supported by the National Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC). DPA is funded by a National Institute for Health Research Senior Research Fellowship. TDS is funded by the Department of Health of the Generalitat de Catalunya under the Strategic Plan for Research and Innovation in Health (PERIS; SLT002/16/00308). RK was supported by the European Social Fund via the IT Academy programme. SR was funded by Estonian Research Council grants PRG1095 and RITA1/02-96-11. The views expressed in this publication are those of the authors and not those of their respective institutions. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit it for publication.

Declaration of interest

CY, RDW, JRA, EB, TDS, MJ, RK, JAK, LK, APU, SR, HMS, and COT report no competing interests related to this work. JNS, AGS, JW, and PR are employees of Janssen Research & Development, a pharmaceutical company of Johnson & Johnson, and shareholders of Johnson & Johnson. At the time the study was conducted, ESB was an employee of Janssen Research & Development, a pharmaceutical company of Johnson & Johnson, shareholder of Johnson & Johnson, and shareholder of Takeda Pharmaceuticals. LC reports her institute has been hired for methodological consultancy by AbbVie Spain, S.L.U., Astellas Pharma, SA, Bristol-Myers Squibb, S.A.U. (BMS), Daiichi-Sankyo España, S.A., Dentsply Sirona Iberia, S.A.U., Eisai Farmacéutica, SA, Fresenius Kabi España, S. A. U., Laboratorios Gebro Pharma, SA, Lilly, S.A., Merck Sharp & Dohme España, S.A., Novartis Farmaceutica, SA, Pfizer, S.L.U., Roche Farma, S.A, Sanofi Aventis, UCB Pharma, S.A., outside the submitted work. KC reports consultancy fees from Eli Lilly, AbbVie, and Pfizer, outside the submitted work. WF is an employee and shareholder of Eli Lilly. AHO is an employee of Lilly Deutschland GmbH. JL reports grants from Versus Arthritis, grants from Medical Research Council, outside the submitted work. AM is an employee of Bayer AG. At the time the study was conducted, KM was an employee of Servier. KV works for a research institute which receives/received unconditional research grants from Yamanouchi, Pfizer/Boehringer Ingelheim, Novartis, GSK, UCB, Amgen, Chiesi, none of these are related to the content of this paper. DV reports personal fees from Bayer, during the conduct of the study and outside the submitted work. DPA reports grants and other (DPA's department has received fees for speaker services and advisory board membership) from AMGEN, grants, non-financial support and other (DPA's department has received fees for consultancy services) from UCB Biopharma, grants from Les Laboratoires Servier, outside the submitted work; and Janssen, on behalf of IMI-funded EHDEN and EMIF consortiums, and Synapse Management Partners have supported training programmes organised by DPA's department and open for external participants. PRR works for a research institute who receives/received unconditional research grants from Yamanouchi, Pfizer-Boehringer Ingelheim, GSK, Amgen, UCB, Novartis, AstraZeneca, Chiesi, Janssen Research and Development, none of which relate to the content of this work.

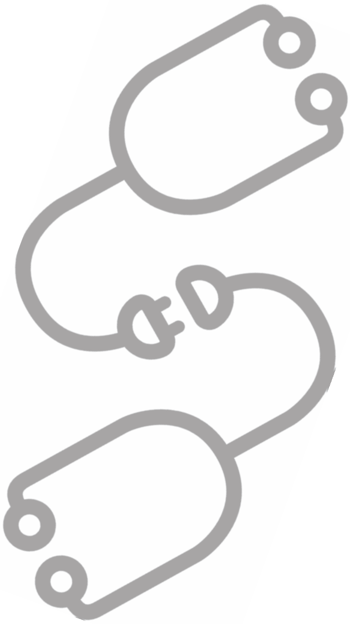
Data sharing statement

The data underlying this article are available at https://data.ohdsi.org/EhdenRaPrediction/.

BIBLIOGRAPHY

1. Listing J, Gerhold K, Zink A. The risk of infections associated with rheumatoid arthritis, with its comorbidity and treatment. Rheumatology (Oxford). 2013;52(1):53-61.
2. Peters MJ, Symmons DP, McCarey D, Dijkmans BA, Nicola P, Kvien TK, et al. EULAR evidence-based recommendations for cardiovascular risk management in patients with rheumatoid arthritis and other forms of inflammatory arthritis. Ann Rheum Dis. 2010;69(2):325-31.
3. Turesson C, Matteson EL. Malignancy as a comorbidity in rheumatic diseases. Rheumatology (Oxford). 2013;52(1):5-14.
4. Dougados M, Soubrier M, Antunez A, Balint P, Balsa A, Buch MH, et al. Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA). Ann Rheum Dis. 2014;73(1):62-8.
5. Turesson C. Comorbidity in rheumatoid arthritis. Swiss Med Wkly. 2016;146:w14290.
6. Singh JA, Saag KG, Bridges SL, Jr., Akl EA, Bannuru RR, Sullivan MC, et al. 2015 American College of Rheumatology Guideline for the Treatment of Rheumatoid Arthritis. Arthritis Rheumatol. 2016;68(1):1-26.
7. Smolen JS, Landewe RBM, Bijlsma JWJ, Burmester GR, Dougados M, Kerschbaumer A, et al. EULAR recommendations for the management of rheumatoid arthritis with synthetic and biological disease-modifying antirheumatic drugs: 2019 update. Ann Rheum Dis. 2020.
8. Arts EE, Popa CD, Den Broeder AA, Donders R, Sandoo A, Toms T, et al. Prediction of cardiovascular risk in rheumatoid arthritis: performance of original and adapted SCORE algorithms. Ann Rheum Dis. 2016;75(4):674-80.
9. Crowson CS, Hoganson DD, Fitz-Gibbon PD, Matteson EL. Development and validation of a risk score for serious infection in patients with rheumatoid arthritis. Arthritis Rheum. 2012;64(9):2847-55.
10. Crowson CS, Rollefstad S, Kitas GD, van Riel PL, Gabriel SE, Semb AG, et al. Challenges of developing a cardiovascular risk calculator for patients with rheumatoid arthritis. PLoS One. 2017;12(3):e0174656.
11. Curtis JR, Xie F, Chen L, Muntner P, Grijalva CG, Spettell C, et al. Use of a disease risk score to compare serious infections associated with anti-tumor necrosis factor therapy among high- versus lower-risk rheumatoid arthritis patients. Arthritis Care Res (Hoboken). 2012;64(10):1480-9.
12. Curtis JR, Xie F, Crowson CS, Sasso EH, Hitraya E, Chin CL, et al. Derivation and internal validation of a multi-biomarker-based cardiovascular disease risk prediction score for rheumatoid arthritis patients. Arthritis Res Ther. 2020;22(1):282.
13. Solomon DH, Greenberg J, Curtis JR, Liu M, Farkouh ME, Tsao P, et al. Derivation and internal validation of an expanded cardiovascular risk prediction score for rheumatoid arthritis: a Consortium of Rheumatology Researchers of North America Registry Study. Arthritis Rheumatol. 2015;67(8):1995-2003.
14. Strangfeld A, Eveslage M, Schneider M, Bergerhausen HJ, Klopsch T, Zink A, et al. Treatment benefit or survival of the fittest: what drives the time-dependent decrease in serious infection rates under TNF inhibition and what does this imply for the individual patient? Ann Rheum Dis. 2011;70(11):1914-20.
15. Wang D, Yeo AL, Dendle C, Morton S, Morand E, Leech M. Severe infections remain common in a real-world rheumatoid arthritis cohort: A simple clinical model to predict infection risk. Eur J Rheumatol. 2020.
16. Jani M, Barton A, Hyrich K. Prediction of infection risk in rheumatoid arthritis patients treated with biologics: are we any closer to risk stratification? Curr Opin Rheumatol. 2019;31(3):285-92.
17. Van Calster B, Wynants L, Timmerman D, Steyerberg EW, Collins GS. Predictive analytics in health care: how can we know it works? J Am Med Inform Assoc. 2019;26(12):1651-4.

18. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-75.
19. Khalid S, Yang C, Blacketer C, Duarte-Salles T, Fernández-Bertolín S, Kim C, et al. A standardized analytics pipeline for reliable and rapid development and validation of prediction models using observational health data. Computer Methods and Programs in Biomedicine. 2021;211:106394.
20. Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. J Am Med Inform Assoc. 2012;19(1):54-60.
21. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
22. Moons KGM, Altman DG, Reitsma JB, Ioannidis JPA, Macaskill P, Steyerberg EW, et al. Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): explanation and elaboration. Ann Intern Med. 2015;162(1):W1-73.
23. Charlson ME, Pompei P, Ales KL, MacKenzie CR. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. J Chronic Dis. 1987;40(5):373-83.
24. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.
25. Young BA, Lin E, Von Korff M, Simon G, Ciechanowski P, Ludman EJ, et al. Diabetes complications severity index and risk of mortality, hospitalization, and healthcare utilization. Am J Manag Care. 2008;14(1):15-23.
26. Romano P, Roos L, Jollis J. Adapting a clinical comorbidity index for use with ICD-9-CM administrative data: differing perspectives. J Clin Epidemiol. 1993;46(10):1075-90.
27. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996;58(1):267-88.
28. Collins GS, Ogundimu EO, Altman DG. Sample size considerations for the external validation of a multivariable prognostic model: a resampling study. Stat Med. 2016;35(2):214-26.
29. Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW, Topic Group 'Evaluating diagnostic t, et al. Calibration: the Achilles heel of predictive analytics. BMC Med. 2019;17(1):230.
30. Ramspek CL, Steyerberg EW, Riley RD, Rosendaal FR, Dekkers OM, Dekker FW, et al. Prediction or causality? A scoping review of their conflation within current observational research. European journal of epidemiology. 2021:1-10.



General Discussion

Ross D. Williams
Jenna M. Reps
Jan A. Kors
Peter R. Rijnbeek

Why do we want to make predictions in healthcare? It is obvious that if we could predict perfectly what would happen to our patients, we could implement proper treatment plans to try to intervene in time and avoid a potential bad outcome. Clearly, this perfect world does not exist, but nevertheless the health care providers (HCPs) are asked for predictions on a daily basis by their patients. When a patient visits a HCP they seek an answer to questions like "What is my chance of heart failure now I have been given the diagnosis of diabetes considering my medical history?". But what options do the HCPs have to answer these kind of prediction questions? First, they could refer to an estimate obtained by descriptive epidemiologic studies that characterise patients at the population level. For example, on average the chance of patients with diabetes to get heart failure is X%. This does not really answer the personalized question the patient has asked but it is often the only available option. Second, the HCP could try to compare this patient with other patients they have encountered before who have a comparable patient profile. This is a difficult and often impossible task considering the complexity of diseases and low amount of available data to support this.

A better option would be to automatically learn from data and develop a prediction model to support the HCP. A prerequisite for building and validating such prediction models is a large amount of such data. Fortunately, data available from EHRs, claims and registry databases around the world is massive. This creates big opportunities to use data-driven methods to improve patient care. A so-called learning healthcare system in which we can use each individual data point generated by a patient to improve the life of a next patient, slowly becomes a reality. However, important building blocks are needed to achieve this ultimate goal at the necessary scale.

Firstly, we have to solve the problem of limited interoperability of healthcare data. This refers to semantic interoperability challenges since the databases use different coding systems (ICD9, ICD10, ICPC, SNOMED, etc.) and syntactic interoperability problems due to different database structures. This makes it nearly impossible to perform studies using standardised analytics in a network of databases. As shown in this thesis, the use of the OMOP-CDM including its incorporated standardised vocabularies has a high impact in unlocking data to improve patient care. Its strong uptake across the world enables observational research at an unprecedented global scale. For example, the European Health Data and Evidence Network (EHDEN, www.ehden.eu) project is standardising more that 166 databases to the OMOP-CDM, including general practitioners databases, claims, registries etc (https://www.ehden.eu/datapartners/) .

Secondly, we need to have established best practices, and standardised tools, to generate reliable evidence from this large pool of standardised data. OHDSI has been on the cutting edge of research to develop standardised analytics for characterisation, population-level effect estimation, and patient-level prediction, and has demonstrated its use in many global studies. Relevant for this thesis is work done on creating a framework for the development and validation of patient-level prediction (PLP) models (1, 2). These articles together provide a robust framework for the development, validation and external validation of prediction models. By following the principles laid out in these papers, researchers can develop models that have a robust body of reliable evidence.

Generating Reliable Evidence

What do we mean with reliable evidence? To generate reliable evidence, the model development and validation should be repeatable, reproducible, and replicable. To have clinical impact the prediction model should have a high discriminative performance and should be well calibrated. Its performance should be generalizable through assessment in a network of diverse types of databases, and the model should be robust (see Table 1):

- **Repeatable** evidence is obtained when all elements, namely the question, researcher, data, and analysis are held constant, and the same result is produced. This seems obvious and easy to obtain but proves to be challenging when common analytic code cannot be applied. Big steps are made by splitting the journey from source data to reliable evidence into two components: standardisation to the OMOP-CDM, and PLP framework code base.
- **Reproducible** evidence produces identical results when all elements are constant except the researcher. This requires that all the steps in the process are fully documented and all analytical steps are performed fully automatically. Clearly, also here the PLP framework enables this.
- **Replicable** evidence combines the same question and analysis with similar data to achieve similar results. For this it crucial to understand whether differences in the results are only due to the differences in data and not because of the implementation of the analysis. For example, if we share a protocol or even a statistical analysis plan, but not analysis code, with data partners in our network, we cannot be sure they implement the analysis in the same way. For PLP this is facilitated by performing development and validation on similar datasets standardised to the OMOP-CDM using common methods.
- **Generalizable** evidence takes the same question and analysis and produces similar results on different data. For example, if a prediction model transports well to another setting this would strengthen our belief that this model is reliable.
- **Robust** evidence is achieved when the same question approached with a different analysis gives similar results. This desired attribute can only be obtained if we have tools that can be parameterized and can be run on a large number of databases. If the conclusion of the study

Table 1. Desired attributes for reliable evidence.

Desired Attribute	Question	Researcher	Data	Analysis		Result
Repeatable	Identical	Identical	Identical	Identical	=	Identical
Reproducible	Identical	Different	Identical	Identical	=	Identical
Replicable	Identical	Same or Different	Similar	Identical	=	Similar
Generalisable	Identical	Same or Different	Different	Identical	=	Similar
Robust	Identical	Same or Different	Same or Different	Different	=	Similar

is consistent over different analysis settings this would improve our trust in the generated evidence. For PLP, this includes running different kinds of algorithms or changes in the parameters such as the lookback period or the time-at-risk window length.

Part 1 of this thesis consists of multiple methodological papers that aim to further improve the best practices for generating reliable evidence for PLP by contributing to model development, validation, and dissemination. In Part 2, two clinical applications of prediction models have been presented, which incorporate some of the best practices discussed in Part 1. In the next sections we will discuss all the contributions and will then identify opportunities for future research.

PART 1 – PATIENT-LEVEL PREDICTION MODELLING

The aim of our work is to develop and enforce best practices for PLP modelling by making available a standardised framework for development, validation, and dissemination. In the sections below the contributions of this thesis are discussed for these three components.

Model Development

When attempting to improve the performance of a prediction model, different algorithms can be applied, e.g. logistic regression, random forests, deep learning, etc. These models are often trained on a subset of the data from a single database. This approach often impacts the generalizability of the model since the data used for training and internal validation may be different in another setting. This can be due to unique elements of the training database which then do not generalise outside of this setting. As an example, the IPCI database is a Dutch primary care database and as such contains a subset of all primary care patients in the Netherlands (3). Whilst it is hoped that this will provide a representative subset of Dutch patients in primary care, it will almost certainly not provide a representative sample of patients in the same setting in other countries. For example, a British primary care database or a US claims database represent a different healthcare system and patient mix. This heterogeneity presents an issue for generalisability, and is not solved through the standardisation to the OMOP CDM. However, we could still take this heterogeneity into account from a prediction modelling perspective. The first option would be to pool the data from diverse sources into one dataset and train a model using this data. Whilst an interesting idea, this is infeasible due to patient privacy concerns preventing the sharing and pooling of patient-level data. The second option would be to use something termed "distributed learning". This is the use of a single algorithm that can be trained at multiple sites simultaneously, to produce a single algorithm. This removes the need for sharing data but does mean that there needs to be connectivity between the databases which can be similarly problematic to the sharing of data. There exist some promising so-called "one-shot" algorithms that provide this distributed learning whilst requiring only one instance of a connection, but the administrative burden of this method is comparatively high. The third option, discussed in Chapter 1 of this thesis, is to use ensemble learning to combine models developed in multiple databases. The traditional methodology behind building an ensemble model is to use multiple different algorithms on the same data source. The principle behind this is that different algorithms are better at dealing with different patterns in the data, and that by using different algorithms the increase in heterogeneity of the search strategy leads to better performing models as they are more adaptable. Using different algorithms is not the only way to increase this heterogeneity. Instead of changing the algorithm, it is also possible to change the underlying data. Considering the context of a federated network, this provides a clear opportunity to utilise the heterogeneity of the different data sources to potentially improve the performance of the model overall.

Learning patient-level prediction models across multiple healthcare databases

Building an ensemble using models developed in a federated network only requires researchers to develop a model in their own database and then share the model they developed. This method of using a network to learn an ensemble model does not require any additional steps to what is done when developing an ensemble prediction model normally. After the sharing of the developed models, they can be combined to then produce a single output. There are a multitude of ways to combine, from a simple majority voting up to the complexity of training a new model using the outputs of the base models. What we learnt from the experiment described in Chapter 1, is that ensemble models generally return an improvement in model performance when compared to the base models, and that the ensemble models tend to transport better as well, when considered in terms of discrimination performance. Calibration was a problem when externally validating both base and ensemble models, but this can be corrected through re-calibration. Another finding from the study was that the choice of ensemble method, also termed the "ensemble heuristic", was important for the transportability of the model. The ensemble heuristic is an important part of the design process when developing an ensemble model. The findings of this study show that using diversified data can aid in producing more robust, better performing prediction models. An important takeaway is that despite the performance increases, there is a cost associated with ensemble modelling. Firstly, there is the added complexity of the modelling process which involves multiple databases and thus likely multiple researchers collaborating across institutions to develop models for model development. Secondly, the use of ensemble techniques necessarily drives up the complexity of the model and this can then act as a barrier to implementation. The heuristic chosen can also increase the complexity of both the model and the implementation process as some heuristics require labelled data to be sampled from the new database in order to be applied. The final models in this study often contained hundreds of covariates that were combined in complex ways and this impacts the implementation of the full model in practice.

When developing prediction models from observational data, a common issue that can arise is that there is a lack of adequate data to perform the required development and validation steps. This can be because the disease in question is relatively rare and as such any single database is unlikely to contain enough patients for model development, or because it is a novel disease (e.g. Covid-19) and as such the records of the patients have not yet filtered through into the research databases. This presents a challenge to both the development and the validation of the models. The development usually involves splitting the dataset and reserving some of the data for the validation. When data is scarce, this splitting can present an issue. There are a few ways that this scarcity can be tackled. The first is to wait until more data is available, although in a time-sensitive situation such as in a novel pandemic this is a costly sacrifice. A way of immediately using available data, and as such not having to sacrifice the speed of development, is to use a proxy disease. A proxy disease is a disease that is used for training of the model to then be applied in a different disease cohort. The central idea behind the proxy disease is that the same patients who are vulnerable to the target disease are also vulnerable to the proxy disease. Due to this common vulnerability, if a model can identify the patients for the proxy, the performance should transfer over to the target disease. Importantly, by not using any data from the target disease cohort for training, all of the target disease data is then kept for the validation of the model in the target disease. This then provides a stronger body of evidence for the performance of the model in the target setting due to the increased cohort size used in the validation.

The research described in Chapter 1, demonstrates the utility of using a proxy disease to develop a parsimonious prediction model. The three COVER models with 9 predictors that were developed using influenza data perform well for COVID-19 patients for predicting hospitalization, intensive services, and fatality. The scores show good discriminatory performance, which transferred well to the COVID-19 population.

From a methodological perspective, this study demonstrates the possibility of transferring performance of models between different patient group settings. Indeed, given the now well understood differences between influenza and coronavirus, the fact that performance is maintained suggests that diseases could be quite different and still vulnerable patients can be effectively identified. This suggests that the developed model can be adopted for other situations where data is scarce, either due to the novel nature of the disease or due to its rarity. The development of a model in abundant data removes the need for many considerations that occur in small data sets, the use of bootstrapping, performance instability etc., and as such is advantageous. The evidence produced by this study is that even with known differences in disease presentation and severity, it is still possible to develop a model using the abundant available data with the proviso that there is likely a need for recalibration due to inherent differences in diseases. One potential application of the methods described in this chapter is in the development of prediction models for rare diseases. Previously it was either challenging or impossible to assemble enough data to adequately develop and evaluate a model for a rare disease, but our findings suggest that it may be possible to use a proxy to develop a prediction model for these situations. This remains an open research question of much potential value.

The methods developed in Chapter 1 also demonstrated the implementation of a technique for parsimonious model development. The lower number of covariates in a parsimonious model, while achieving adequate performance, makes it much easier to implement than a data-driven model that includes hundreds of covariates. In this chapter we developed baseline, data-driven and parsimonious models. These serve different purposes and together provide a stronger body of evidence for the use (or rejection) of prediction models in this setting. Firstly, the data-driven model provides an upper bound of performance. This is the model which, in theory, provides the best possible performance in the setting. However, these models often contain hundreds of covariates, and these covariates are also specific codes rather than clinical concepts. The large number of covariates present a clear barrier to clinical implementation of the models. This motivated the development of the baseline models, which represent a lower bound for performance but for a high level of applicability. These models use only age and sex as covariates and as such are very easily implemented. The performance of the baseline and the data-driven models can then be compared. If the performance difference is small, then likely the baseline model can be used clinically. If the performance difference is large, then it demonstrates that an increase in model complexity provides a valuable performance increase. In this scenario the development of a parsimonious model could find the correct balance in the complexity-implementation performance trade-off. In order to develop this, candidate covariates are created and a limited set of these are then used in the model development to create models with a lower number (<20) covariates. After training the parsimonious model, the performance can be compared to the data-driven model, and if there is a large gap the parsimonious modelling can be reapplied, using e.g. different covariates, to attempt to further increase the performance. Once the performance is deemed close enough to the data-driven performance, the model can be deemed ready for implementation.

Model Validation

Proper model validation is needed before implementation in clinical practice. This should include both internal validation and external validation. In order to externally validate a model, it is helpful to have a high level of interoperability between the source and validation data. By taking a collaborative approach to model development and external validation through the widespread adoption of the OMOP CDM the external validation of prediction models is much better facilitated.

External validation in a network

When considering trust in models and personalising treatment, several key points need to be addressed. These are: does my model perform well on the data it is trained on, does it perform well on new data of the same type (e.g. a different EHR or claims database), does the model maintain performance when applied to new data of a different type? In effect the question we need to answer is, can I trust this prediction model to continue to perform at an adequate level when being used to treat patients in a clinical setting? In order to answer this question, models can be externally validated. Historically, external validation has been a slow process, with the average time to external validation of a model being three years or more (4), and many models never receiving any external validation although this does appear to be improving as more attention is paid to the topic (5). The problems of externally validating models are due to issues of model sharing, cohort definition sharing, poor setting of prediction problem, and more generally there is an issue that is more widespread within science of validation studies not being performed. Concerning the technical issues with external validation, the increasing interoperability of healthcare databases, through the widespread adoption of the OMOP CDM, facilitates external validation of prediction models. By having a common semantic and syntactic structure, models can be trained on a database and then can easily be shared between researchers, including all necessary data and cohort definitions, to then be validated in new datasets without the need for a data engineering step, a major step forward in prediction modelling.

We have seen, however, that the removal of technical challenges does not guarantee high performing models. Often when externally validating a model we see performance changes, usually a drop in performance, and this can be concerning when considering whether a model is adequate for clinical use. When considering the use case in chapter 3, in which multiple stroke prediction models were externally validated, we saw that performance frequently decreased from the development setting. This is a common observation for external validation of prediction models. There remain several questions, including why this performance drop occurs, and what it actually tells us about the models.

Should we immediately dismiss a model with poor external validation in one setting? Or should we prohibit its use in the specific setting in which it performed poorly? Given the

Table 2 Common performance assessment metrics and their definitions

Performance Metric	Definition
Area under receiver operator curve	Calculated by plotting an ROC curve (y = true positive rate, x = false positive rate) and then measuring the area under this curve. This is a measure of discrimination that measures how likely for a pair of patients, that the patient who is at higher risk is to be assigned the higher risk of the outcome given by the model.
Area under precision recall curve	Calculated by plotting a precision recall curve (y = precision, x = recall) and then measuring the area under this curve.
Calibration in the large	A measure of calibration which compares the average expected risk to the proportion of the patients that get the outcome.
Calibration Slope	A measure of calibration where the slope of the outcome model is compared to the ideal slope of 0.5.
Graphical calibration (LOESS)	This is a manner of calibration assessment where a local area regression method is fit to the output of a prediction model. This is then compared to a slope with intercept =0 and gradient =0.5.

essential heterogeneity between different databases (for example a Dutch general practice database and hospital database from the US), what does a change in performance between the two settings mean? Externally validating from one to the other gives some information on the generalisability of a model, but not what the performance means in the local context. In order to make meaningful decisions surrounding potential clinical impact we need to give proper context to the performance. To address some of these questions, Chapter 4 discusses a new model development and external validation method called Iterative Pairwise External Validation (IPEV). IPEV utilises a network of databases to develop and validate multiple prediction models on the same problem setting. It trains a baseline and data-driven model at each site and then rotates these models through the databases. Once each model has been validated in every database, we can give the performance an important context by looking at how a model's external validation performance compares to its internal validation performance. For example, what is the difference in performance between a model developed in IPCI validated in Optum Clinformatics, and a model developed in Optum Clinformatics. This comparison allows us to say more about the generalisability of the model and how this compares to the expectations we have for performance in a database. If a model trained in a database achieves low performance, then we would not expect a model externally validated in that database to have high performance.

IPEV is a simple way of assessing performance in context and vastly improves our understanding of the models produced and of the heterogeneity of the databases included. By using IPEV we can say whether there is a meaningful difference in the expectation of model performance between two databases. This provides some explanation of the difference in performance between internal and external validation of a model. Another benefit of the method is to provide context to the impact of complexity on performance. By using this context we can demonstrate that increased model complexity provides a measurable benefit (or indeed show that this is not the case).

In Chapter 4, which examined the application of IPEV to the use case of predicting heart failure in type-2 diabetic patients, one of the databases produced lower performance for all models externally validated on it. These performances were in line with performance of the model developed using the data contained in this database. This tells us that it is harder to make predictions for the population in this database. The question that remains here is, what does this tell us about the performance of the model in general, and how does this performance drop affect our understanding of the model itself? If a model is externally validated using this population, we expect a lower performance. As such, when observing a lower performance here than in either the internal validation or external validations of a prediction model we should not then dismiss the model as low performing. We can continue to use the model in the context it was developed in but we should be aware of the population it does not work in. The knowledge of where it does and does not perform well can be then utilised to inform where we choose to apply the model. For example, if one database contains mostly younger patients and a reduction in performance is seen, it should be seen as a warning about the performance of the model in younger patients and prompt investigation as to whether this is unique to the database or if this is a more commonly seen issue. This is a complex issue without a simple answer. Whilst age is often included in models it often has strong interactions with other covariates and the overall case-mix of a younger group of patients is likely to differ significantly even when controlling for age. Why performance drops is an open, important and interesting research question.

There remain many questions still such as, why do the models perform differently, is it solely due to case-mix differences between databases? However, by combining these ideas of external validation in a network and IPEV, we can go someway to reinforcing the trust that a model has by providing context on its performance, explanation of a performance drop, and reassurance that the increased complexity is necessary for the desired results.

Model Dissemination

The DELPHI library

Developing models and adequately validating them is not, however, the end of the story. Of critical importance to the implementation of PLP models into clinical practice is the dissemination of this research. Currently models are developed, a scientific article is published in a journal, and sometimes the models are made available. The TRIPOD reporting standard goes some way to addressing the inadequacies that many prediction model articles have, but it does not solve the larger problem of dissemination of the model itself. There are some efforts such as the PACE repository from Tufts medical centre(21), which contains a description of the model and often the model itself in a tabular format. A model in a tabular format lacks an essential interoperability that would make it easy to use across multiple sites. The cohort information will need to be translated to different settings, and similarly the covariates themselves will need that. A solution to these problems is found through the widespread adoption of the OMOP CDM. By developing models using data in this common format, and by using a standardised pipeline for the development, a standard results and model output format can be created. This format can then be incorporated into a repository of models that allow for the easy accessing, evaluation and downloading of these models.

The creation of a database, the DELPHI library, to store the standardised output of PLP models contributes to increased visibility, accessibility and investigation of the models by researchers, clinicians and regulators. Whereas previously interested parties would need to perform a literature search of disparate sources and keywords, they are now able to perform a systematic search within a single database to find available models, ready for implementation. Once a model is found, all relevant model information, for example the performance, definitions of target and outcome cohorts (especially relevant for clinical implementation) and importantly the covariates and model specification will be immediately available. By making all of this information available, interoperable and downloadable, it is hoped that the external validation of models will become increasingly common.

The library will also provide an interactive results explorer. This will allow interested parties to assess the models using a variety of metrics, to observe the differences that occur in performance for different thresholds (e.g. stakeholders might apply different costs (both financial and human) to false negatives and positives. The differing costs of these errors will mean that stakeholders may desire to set different risk thresholds. The library will provide the ability to investigate this.

PART 2 – CLINICAL STUDIES

The second part of the thesis focusses on the development of prediction models for specific clinical problems. These studies produced some well-performing patient-level prediction models. A key lesson learnt from these studies was the need for the involvement of a multidisciplinary team in the model development process. The major work for these studies was conducted during so-called study-a-thons. A study-a-thon involves grouping multiple researchers together in a room to intensively work on a question. This created a dynamic environment that allowed for the rapid progression of the work. The multidisciplinary nature of the team meant that questions could be asked, and answered by experts, with almost no time delay. In doing this, discussions could develop rapidly as the clinical partners who had the research questions could be helped and guided in the development of that question by experts in the data and in prediction modelling. Having the team in a single location helped to stimulate further questions and to highlight some of the challenges that remain in implementing the models into clinical practice.

Predicting all-cause mortality following total knee replacement

The first clinical chapter dealt with the prediction of short-term mortality following a total knee replacement. A knee replacement is a safe and effective procedure for the treatment of knee complaints in arthritic patients. There are however risks associated with the procedure. These risks include amongst others infection, need for reoperation, and death. In this chapter we focussed on the prediction of mortality. Whilst the risk of mortality following the surgery is low (between 0.2 -0.4% (7)), mortality is of high impact and importance to patients. Considering the planning of a patient's treatment pathway, knowing the risk of mortality following a surgery can be useful in the shared decision making process. A patient could be reassured by being given a low risk. Another possibility is that a patient has a higher risk and still opts for the surgery. In this scenario, patients are making a well-informed decision based upon the best information they have available. Aside from having adequate performance, other aspects needed to implement a model in practice are that the model is well reported, clearly describes the situation it is designed to be used in, and is usable in a practical sense (e.g. that it is not too time consuming when calculating a prediction for a patient). The feasibility of application is an issue when considering data-driven models. These are often large complex models with hundreds of covariates. Having such a large number of covariates is clearly a barrier to clinical practice as it is too time consuming for a health-care professional to calculate. Reducing the complexity of a model whilst maintaining performance is imperative to the implementation in practice, and motivated the creation of a 12-feature model in this work. The development of the parsimonious model follows a similar procedure to that discussed extensively in Chapter 1.

A simple prediction model based on sex, age, and 10 comorbidities that can identify patients at high risk of short-term mortality following TKR was developed that demonstrated good, generalisable performance. The 12-feature mortality model is easily implemented and the performance suggests it could be used to inform evidence-based shared decision making prior to surgery and targeting prophylaxis for those at high risk. This study demonstrated the method of parsimonisation of a model. It also provided a use case in leveraging an existing network for external validation to produce a model according to best practices detailed in Chapter 3.

Predicting adverse health outcomes in rheumatoid arthritis

In the second clinical chapter, models to identify which rheumatoid arthritis (RA) patients are at high risk of adverse health outcomes, were developed. Identifying these patients poses a major challenge to the treatment of RA. Being able to identify these patients could provide a major benefit by allowing for more personalised treatment choices, more targeted interventions and in general providing patients with reassurance and helping them to make informed choices on their treatment as part of a shared decision making treatment structure. We aimed to develop and validate prediction models for a variety of adverse health outcomes in RA patients initiating first-line methotrexate (MTX) monotherapy. Final models for serious infection, MI, and stroke demonstrated good performance across multiple databases and can be studied for clinical use. This work again followed best practises for model development and external validation in a network, however it did not include the parsimonisation of the models and as such there is a barrier to clinical implementation unless this, or some form of embedding in the EHR is created. The lack of parsimonisation here was in part due to the number of models developed and the difficulty with the current time and resource intensive development of parsimonious prediction models. This opens the question of how to best automate this process in the future. A perspective of this is given in the following section.

FUTURE WORK

Patient-level Prediction

This thesis addresses some of the issues that impact the widespread adoption of prediction models in clinical practice. In part 1, we discussed different ideas of model development and validation, with a particular focus on the evaluation of models in context and how this can be used to improve the trustworthiness of the models. There remain however many interesting topics for future research as presented in this section.

Model parsimonisation

The current major issue preventing model implementation, is the complexity of the models developed. As has been explored in multiple chapters in this thesis, a simpler model is preferred over a complex model with little or no performance gain. In Chapters 2 and 6 we have explored the use of a 2-step modelling technique to produce parsimonious models. This parsimonisation process was an effective method for producing simple models; however, it has a couple of limitations. A major limitation is that there is a high demand on clinician expertise in the modelling process. Clinician input is needed for the assessment of the data-driven covariates and the creation of various cohorts to be used as covariates. In Chapter 6 this was feasible as there was only a single model developed, but this process does not scale well to the creation of multiple models, as was experienced during the research phase that produced the content of Chapter 7. The inclusion of this manual step is time consuming, and as such an automatic or semi-automatic process to produce the parsimonious models would be preferable. Fortunately, there exist several ways to produce parsimonious models using automated feature selection. A simple option is to increase the penalisation in the LASSO method to such an extent that very few covariates are included in the model, but given the L1 properties of this algorithm, LASSO with high penalisation is prone to overfitting and the performance collapses when externally validated. Work which is currently ongoing, involves experiments with the iterative hard thresholding (IHT) algorithm. This uses an L0 penalisation parameter that selects the best *k* variables to approximate the *true* function that produces the data. *k* here is a hyperparameter that decides where the hard threshold of (IHT) is fixed, typically 5, 10 or 15 variables. This procedure has been shown in preliminary research to better maintain performance when externally validated than a LASSO model. This work will be extended to include more parsimonisation methods and be applied to more problem settings to produce a more robust analysis of the performance of these different parsimonisation processes. The goal of this is to discover a fast, effective and scalable parsimonisation method which would allow for the development of more prediction models that have a greater chance of clinical impact.

Multi-database learning

In Chapter 1 we considered how using multiple database to *ensemble* a prediction model would affect the performance and we saw that the ensemble did give improvements in discrimination performance but would likely need to be recalibrated. We built an ensemble of linear regression models, each model built in a separate database, but another option is to ensemble different algorithm types in the same data. Within the PLP package, aside from LASSO regression, there are also random forest, gradient boosting machines, deep learning methods etc. available. An option to utilise this diversity of available methods is to train a selection of these models and then ensemble them (there are also multiple options for this as explored in chapter 1). Currently we are training ensembles of random forest, gradient boosting machine, and LASSO in a US claims database with the intention to externally validate an ensemble of these models and observe how this model performs and if it sufficiently increases the model performance to warrant the increase in complexity. One repeated finding in experiments with ensemble modelling is that the calibration of the final models is often poor. This occurs both for ensembles developed using 1 database as well as in multi-database ensembles. Potential reasons for this are the differences in background event rates between the training and validation settings and the fact the models are often weighted on discrimination and not calibration, meaning calibration is not considered in the ensembling process. Methods of calibration assessment and recalibration will be important in ensemble models and are to be further developed for clinical use.

Calibration and recalibration

Almost every study in this thesis used calibration at some point as a performance measure. Calibration is however an often neglected part of the prediction model assessment. The impact of calibration on clinical performance has been demonstrated but the potential for recalibration of models is yet to be fully explored. This can be split into 3 different broad categories. The first of these is to simply adjust the intercept of a model, this is the simplest and crudest form of recalibration but has produced good results. When validating a model in a new database, if the model consistently underestimates risk then simply increasing the value of all risks by the same amount can show improvement in the calibration (preserving the discrimination performance). The second method is to adjust the intercept and the slope. This is useful when the amount of miscalibration varies across the range of predictions and if there is a linear relationship between the risk size and miscalibration size. It can be more effective in correcting this than a simple intercept adjustment. The third and more complex version of recalibration (and indeed there is a question as to whether this can be considered a recalibration) is to use all the covariates selected by the original model and then to refit the coefficients using new data. This will likely have impact on the discrimination performance but should improve the calibration of the model. There is an epistemological question here of whether this is a recalibration or simply developing a new model and as such consideration needs to be made if the evidence from the internal validation of the original model can still be considered as evidence for the new model or if the *recalibrated* model should simply be assessed on the new evidence. What all of these methods require, however, is new data in order to be able to recalibrate and they are all done *post-hoc*. That is to say the model has been run originally and then the adjustment is based upon the errors seen in the model. I would like to investigate if it is possible to perform beneficial recalibration without the need to use any post-hoc assessment and adjustment. An ideal situation would be to use some existing knowledge (e.g. using summary statistics surrounding differences in event rate and prevalence between original and new data source) to be able to recalibrate a model effectively.

Model Fairness

Another major focus of future prediction research will be on social determinants of health. Recently there has been a major focus on the provision of adequate and fair healthcare to all sectors and peoples in society. It has been demonstrated that historical medical practice has unfairly ignored or harmed people from vulnerable and minority groups. A major reason for the unfairness is the lack of representation of these people in the data used for conducting research. If a patient population is insufficiently represented in the data used for analysis, then the model or analysis could perform poorly in this subgroup, but this is then not seen in the general results. This raises the question of how to best conduct the assessment of performance variation between people from various socio-economic and ethnic backgrounds. In terms of prediction modelling, the first task is to assess if the models developed within the OHDSI PLP framework have different performances for different demographic subgroups. Given that it is likely differences will be measured, the next task is to figure out how to adjust for this. One common issue is poor calibration in the subgroup. In theory, this could be fixed using recalibration. If we observe a systematic over or under estimation of risk, then a simple baseline risk adjustment can help. If the miscalibration is more complex, methods such as slope updating and covariate refitting could be used. These ideas were discussed earlier, in the recalibration section, and the same logic as for the entire population applies here when thinking of adjustment for a subgroup. In assessing the differences in performance, simply analysing the performance in the subgroup using standard metrics can present an incomplete picture. Specialised metrics such as equalised odds can provide a more complete picture and in fact could be used in the training process to develop models without fairness issues in the first place.

Of particular importance to this process is to be mindful at every step that research in this area can have negative as well as positive impacts on the healthcare of the groups that have traditionally been ignored or underserved by the healthcare establishment. As such it is essential to include the voices of the affected people in the conduct of this research.

Finally, what have we learnt along the way? Firstly that heterogeneity of patients can be used to our advantage. The individuality can be used to personalise treatments to maximise the benefit. This can be done through the use of risk-guided intervention choices that are supported by PLP models. Secondly, in order to do this we need to have robust, reliable, reproducible and contextualised evidence. In order to produce this, open source and open science principles must be applied at all points throughout the research and implementation process. Finally, models need to be shared, reported, tested, criticised and improved by a community of researchers with a common goal of using massive healthcare data to create evidence to improve the treatment pathways of patients. In a learning healthcare system we have a responsibility to the patients who have provided the data, to use that data. This thesis detailed multiple ways to best leverage the information contained in this data to improve the patient pathways. Prediction modelling has the potential to transform treatment decision making. Simple, understandable and applicable models are needed to bring about this revolution in care.

BIBLIOGRAPHY

1. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018.
2. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
3. de Ridder MAJ, de Wilde M, de Ben C, Leyba AR, Mosseveld BMT, Verhamme KMC, et al. Data Resource Profile: The Integrated Primary Care Information (IPCI) database, The Netherlands. Int J Epidemiol. 2022.
4. Siontis GC, Tzoulaki I, Castaldi PJ, Ioannidis JP. External validation of new risk prediction models is infrequent and reveals worse prognostic discrimination. J Clin Epidemiol. 2015;68(1):25-34.
5. Ramspek CL, Jager KJ, Dekker FW, Zoccali C, van Diepen M. External validation of prognostic models: what, why, how, when and where? Clin Kidney J. 2021;14(1):49-58.
6. Wessler BS, Paulus J, Lundquist CM, Ajlan M, Natto Z, Janes WA, et al. Tufts PACE Clinical Predictive Model Registry: update 1990 through 2015. Diagn Progn Res. 2017;1:20.
7. Berstock JR, Beswick AD, Lopez-Lopez JA, Whitehouse MR, Blom AW. Mortality After Total Knee Arthroplasty: A Systematic Review of Incidence, Temporal Trends, and Risk Factors. J Bone Joint Surg Am. 2018;100(12):1064-70.

SUMMARY

At the heart of medicine, lies decision making. Every day clinicians face the complex challenge of making decisions about a patient's treatments. E.g., if a patient has hypertension, which treatment should they choose(1). When deciding on how to treat a patient, the risks of a treatment are compared with the benefits in order to choose an optimum treatment pathway(2). For example, when a patient is already at a high risk of heart failure, adding a drug which is known to increase the risk of heart failure is likely to be discouraged(3-6). Knowing that a treatment increases or decreases the risk of an outcome is not the only relevant element of risk. For any patient an increase in risk of one outcome can be much more impactful than the increase in risk of another. This is the trade-off between relative and absolute risk increases. A small increase in a large absolute risk is likely a bigger, more worrying increase than a big increase in a small risk(7). Frequently, guidelines recommend that treatments diverge based upon the risk of certain outcomes and the different known risk profiles of medications(8, 9). An example of this was seen in chapter 4, where, in the relevant guidelines, the selection of a diabetes medication depends upon risk of heart failure. However, the same guidelines did not include how a patient's risk should be calculated. When applying the guidelines, the doctor must assess what the risk of some event is for a patient. To make this assessment, they might use information available to them such as causal inference studies, experience and clinical intuition. The use of causal inference studies only takes them so far. These studies provide difference in absolute risk, expressed as relative risk across an entire population. This average risk does not account for the heterogeneity seen at the patient level. Furthermore, a clinician's experience is limited as they may not see enough patients to identify rare patterns and intuition is limited by things they do not know, e.g., the patient's full history. As such, the assessment of risk recommended by the guidelines is a very difficult task at the patient level. This severely limits the possibility to personalise the treatment of a patient. In general, the heterogeneity seen at the patient level is considered problematic for treatment(10-14). This same heterogeneity can however be used for personalisation, if we can leverage it in prediction models. By providing a high performing risk model to assess the probability of an outcome, better informed healthcare decisions can be made and ultimately lead to the better personalisation of a patient's medical journey. This personalisation will contribute to a better performing, more efficient healthcare system and reduce the burden on the clinician and patient.

To create better risk models, we need to have standardised data, standardised analytic pipelines and standardised research practices(15). This allows for more direct comparison of models and for the overall increase in trust of the robustness of the methodology used. The best way of implementing these standardisations is to start with the data. If the data we collect for observational research can be mapped to improve the semantic and syntactic interoperability, further standardisations of analytical pipelines become easier. One method is to use the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM)

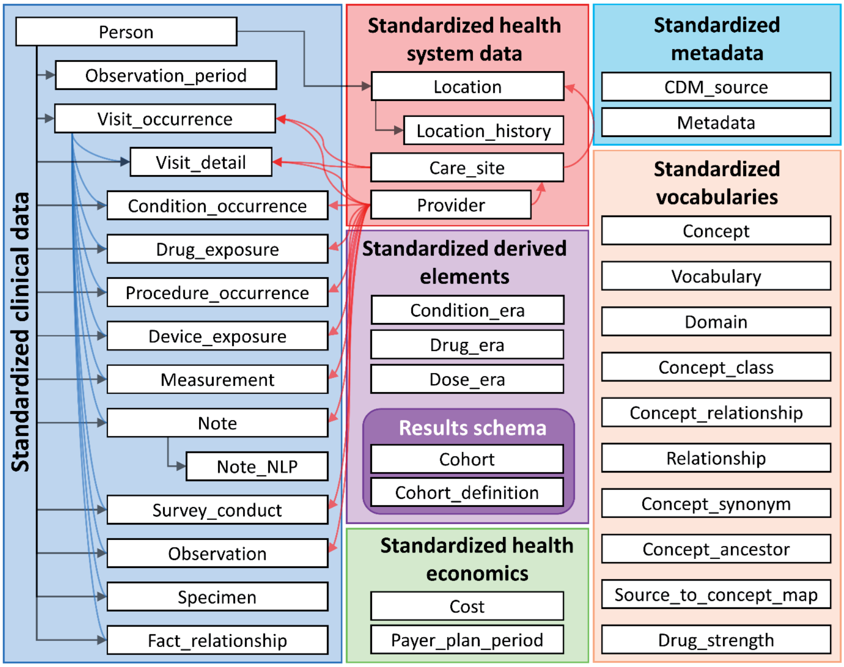


Figure 1 The structure of the OMOP-CDM

maintained by the Observational Health Data Sciences and Informatics (OHDSI) community (see Figure 1)(16).

The OMOP-CDM aims to improve both the syntactic and semantic interoperability of observational health data. Standardising clinical data to a common format as shown in Figure 1 (blue box), enables the use of standardised analytic pipelines such as the Patient-Level Prediction (PLP) framework. The use of standardised vocabularies (orange box) improves the semantic interoperability, i.e., it facilitates the identification of clinical concepts using a common terminology. More information about the OMOP-CDM can be found in The Book of OHDSI (https://book.ohdsi.org).

This thesis details work which helps define best practices in conducting PLP research using data from a federated network of standardised observational databases.

Part I

In Part I we considered how best to develop and evaluate PLP models within federated data networks. A federated network is a network of databases that protect patient information by using a privacy-by-design approach(17). As much health data is siloed and not allowed to be distributed outside of a specific environment, pooling of data is impossible. A federated network removes many of the issues with the siloed of data by implementing a common data model across the network to ensure the databases have a high level of interoperability. As we know the underlying syntactic and semantic database structure, we can share standard analytic tools between databases rather than passing the database to the tool. This means only software and summary statistics are shared and no patient data. All the studies in this thesis were conducted using data mapped to the OMOP CDM. Without this mapping, this work would have been next to impossible.

Chapter 1 assessed whether transportability of models can be improved by ensembling multiple models, each developed in a separate database, and applying this to a new database(18, 19). Here transportability is defined as a model's ability to maintain its performance when being assessed in a new database (20, 21). Any single database often contains a population that is non-representative of the broader disease specific population (e.g., the IBM Marketscan Medicare supplemental database contains older patients who are more affluent). We investigated whether we could combine models developed in different databases (proxy for types of populations) to create a model that is better for the general population (and therefore should transport better). For each question in a set of 21 prediction questions, we trained five single database models each using a different observational healthcare database. We then developed and investigated several different ensemble models that combine, or *ensembled,* the five different models. These ensemble models used the performance of the base (or level 1) models and then applied one of several fusion methods, by matching the new patient to the underlying model based on age, or one of 3 different stacking methods using incremental amounts of data. The stacking methodology fits a logistic regression model to the outputs of the base models to combine these in a more elegant manner (22). This form of ensembling requires data from the validation set. Performance of each model was investigated via discrimination and calibration using data from a new database not used in the model development. The internal validation of a model developed using the hold out database was calculated and presented as the 'hypothetical optimum' for comparison. Fusion ensembles generally outperformed the single database models and were more consistent when applied to new data. Stacking ensembles performed poorly in terms of discrimination when the additional data needed to perform the stacking process was limited. Calibration was poor when ensembles and single database models were applied to new databases. Comparing all the methods detailed above, we observed that in general ensemble methods improve performance over base models. This performance must be considered with the context of vastly increased model complexity. All the ensemble models would need recalibration before clinical implementation.

Chapter 2 considered the use of proxy learning and parsimonisation (23, 24). Proxy learning is the development of a model in a similar 'proxy' population to the main target population of interest. This model is then evaluated in the true problem setting in a new dataset. This is done because there may be, for a variety of reasons, insufficient data to develop in the true population of interest (25). In chapter 2, this involved using data from a different target disease. The standard form of external validation is to perform this validation on an identical problem setting but then in a new database. This chapter details a model developed at the start of the Coronavirus (covid-19) pandemic. As such the models needed to be developed rapidly and at a time when there was limited data available on covid-19 infections. The aim was to develop a usable model as quickly as possible. The parsimonisation involved a 2-step process of first developing a data-driven model, then using clinical expertise to refine the covariates into a more manageable number and then comparing performances. The proxy learning used influenza data, which is abundant, to attempt to develop a model rapidly to use for covid-19 patients. Due to the limited amount of covid-19 data, we decided to use this proxy method to preserve the covid data for validation. This increased the strength of evidence of performance without having to wait longer to collect more data. We developed three models assessing hospitalisation, hospitalisation with intensive services or death and fatality. These three endpoints represent different disease severities. The three models performed well both for complex and parsimonious versions in the influenza dataset for which we sampled 150,000 patients. This performance was largely replicated in the multiple international Covid-19 datasets that were at our disposal (n=44,507). This analysis demonstrated that proxy learning can be an efficient and effective technique. The parsimonious nature of the models developed meant that they can easily be used in multiple settings to affect patient care and strategic planning.

Chapter 3 of this thesis establishes a best practice for conducting external validation (26-28). The chapter demonstrates how the standardisation of data to the OMOP CDM helps to facilitate the development of standardised tools. This reduces the burden for externally validating a prediction model in a network of observational databases. To do this work, we took multiple models from the literature, implemented the models into a OMOP CDM compatible format and applied them across six databases. To demonstrate the utility of this pipeline, a use case study was performed using five existing models that predicted incident stroke in atrial fibrillation patients. The five existing models, (ATRIA, CHADS$_2$, CHADS$_2$VASC, Q-Stroke and Framingham) were able to be integrated into the OHDSI framework for patient-level prediction (29-31). They obtained mean c-statistics ranging between 0.57-0.63 across the six databases. This was comparable with other validation studies. The validation network study was run across six datasets within 60 days once the models were replicated. The techniques in this chapter were shown to be an effective way to externally validate models. The speed at which the models could be validated within the network, after being converted to an OMOP CDM compatible format, demonstrates the power of the standardisation of analytic pipelines in aiding in rapid and reliable evidence generation.

In chapter 4 we considered how we can best assess the performance of a model in an external validation setting (32, 33). This chapter introduced the idea of performance in context. In order to make meaningful statements about how performance differs between internal and external validation, it is helpful to have an expectation of performance in a new population (34). This chapter suggested doing this in a twofold manner. The first is to use a baseline model, we suggest a simple age and sex model, which gives context of how impactful increasing model complexity is. If the baseline model has similar performance to the complex model, then we know that the extra complexity does not provide a relevant performance increase. The second is to set expectations of a "full model" by developing a new model in the external validation database to give an understanding of what the expectation of performance in this database is. If a model has an internal validation performance of 0.7 in database A but external validation in database B of 0.6, this looks bad. However, if we know a model developed in database B has an AUC of 0.62, then the drop in performance when externally validated is likely due to case-mix. Using this procedure, we observed that often a database can be thought of as harder or easier to predict in. This variation in difficulty was demonstrated by a use-case of predicting incident heart failure in diabetes patients in the 1 year following initiation of a second diabetes medication. A total of 403,187 patients were included in the study from 5 databases. We developed 5 models which when assessed internally had a discriminative performance, assessed by c-statistic, ranging from 0.73 to 0.81 and acceptable calibration. When externally validating these models in a new database, three models achieved consistent performance and in context often performed similarly to models developed in the database itself. This study provided insight not only into the potential performance of the clinical model, but also into the databases themselves showing some were more difficult to predict in than others. The process of rotating development and validation databases and implementing a baseline model is called iterative pairwise external validation (IPEV). IPEV demonstrates the potential additive value of using more complex models and gives context to model performance in new databases.

Chapter 5 detailed the development and deployment of the DELPHI prediction model library. The main aim of this was to ease the dissemination and evaluation of prediction models that were developed for OMOP CDM data (35). Given the rapid increase in the development of prediction models over the last 10 years, but the lack of improvement in the reporting of models, there was a clear unmet need for a centralised repository to standardise the model dissemination process. The widespread adoption of the OMOP CDM and the standardised analytics made possible using the PatientLevelPrediction R package (36), mean that many prediction models produced now have a common results object. By leveraging this standard results object, models and their performance can be easily formatted to a database. We developed a database backend that will take the information about a model or the external validation of a model and store it. On top of this we developed a Graphical User Interface that allows users to interact with the database in a simple and intuitive way. Importantly this loads a dynamic results exploration environment. This dynamic environment allows for users of the GUI to explore results and change parameters (for example the threshold for a decision) to see how this affects the model performance. The models themselves can also be downloaded, evaluated on new data, and this performance reuploaded as an external validation. It is hoped that this makes external validation easier and thus more commonly performed. By doing this the DELPHI library aims to turn results from a static object, e.g., a journal article, to a dynamic ecosystem of evidence generation. This should increase the scale at which prediction models will be validated. The DELPHI library represents an important step in the implementation of models in clinical practice by improving the flexibility of evaluation by clinicians, regulators and researchers. It is hoped that this leads to an increase in trust in the field of PLP modelling.

Part II

The second part of this thesis moved away from methods development and considers specific clinical questions that were answered with a multi-disciplinary team. The questions came directly form clinicians that saw a clear unmet need for a prediction model in their daily practice.

Chapter 6 detailed the development and validation of a prediction model for 90-day post-operative outcomes following a total knee replacement (TKR). TKR is a safe and cost-effective surgical procedure for treating severe knee osteoarthritis (OA) (37-39). Although complications following surgery are rare, prediction tools could help identify high-risk patients who could be targeted with preventative interventions (40-42). The aim was to develop and validate a simple model to help inform treatment choices. This chapter discusses the development of a prediction model for mortality that was conducted during a study-a-thon in Oxford that produced multiple articles. The main finding of this chapter was the development of a 90-day mortality prediction model. Both a complex and a simple model were developed for this problem setting. The complex model had a c-statistic of 0.78 internally and 0.70 externally. The parsimonious model had a c-statistic of 0.77 internally and 0.71 externally. This demonstrates the parsimonious model is similarly performant as the complex model and as such is preferred. The performance achieved by the parsimonious model suggests that it could be clinically impactful. The use of only 12 variables in the model means it is implementable and could immediately aid in the decision making around surgery.

For chapter 7, we considered a prediction model built for patients with rheumatoid arthritis (RA). Identification of RA patients at elevated risk of experiencing any of several adverse health outcomes remains a major challenge (43, 44). This chapter discusses the development and validation of prediction models for a variety of adverse health outcomes in RA patients initiating first-line methotrexate (MTX) monotherapy. Models were developed and internally validated on 21,547 RA patients and externally validated on 131,928 RA patients. Models for serious infection (AUC: internal 0.74, external ranging from 0.62 to 0.83), MI (AUC: internal 0.76, external ranging from 0.56 to 0.82), and stroke (AUC: internal 0.77, external ranging from 0.63 to 0.95), showed good discrimination and adequate calibration. Models for the other outcomes showed internal discrimination with AUC < 0.65 and were not externally validated. We developed and validated prediction models for a variety of adverse health outcomes in RA patients initiating first-line MTX monotherapy. Final models for serious infection, MI, and stroke demonstrated good performance across multiple databases. These models could potentially help to personalise treatment by, for example, implementing prophylactic antibiotic use, increased monitoring for patients at high risk, and providing reassurance to patients at low risk. These models are candidates to be studied for clinical use.

In conclusion, prediction models can be impactful in aiding decision making in clinical practice and personalising healthcare. However, there remain multiple challenges in the development of effective and implementable models. To warrant the complex task of implementation, the models must demonstrate a strong ability to positively impact patient care. To positively impact patient care must remain as the main goal of research into patient-level prediction modelling.

BIBLIOGRAPHY

1. Kjeldsen S, Feldman RD, Liu LS, Mourad JJ, Chiang CE, Zhang WZ, et al. Updated National and International Hypertension Guidelines: A Review of Current Recommendations. Drugs. 2014;74(17):2033-51.
2. Muhlbacher AC, Juhnke C, Beyer AR, Garner S. Patient-Focused Benefit-Risk Analysis to Inform Regulatory Decisions: The European Union Perspective. Value in Health. 2016;19(6):734-40.
3. Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. Nat Rev Cardiol. 2011;8(1):30-41.
4. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure. Journal of Cardiac Failure. 2022;28(5):E1-E167.
5. Lippi G, Sanchis-Gomar F. Global epidemiology and future trends of heart failure. AME Medical Journal. 2020;5.
6. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Bohm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. European Heart Journal. 2021;42(36):3599-726.
7. Soares AA, Peto R. The big causes of death from noncommunicable disease. B World Health Organ. 2016;94(6):413-4.
8. Care F. Standards of Medical Care in Diabetes 2019. Diabetes Care. 2019;42(Suppl 1):S124-S38.
9. Yu T, Vollenweider D, Varadhan R, Li TJ, Boyd C, Puhan MA. Support of personalized medicine through risk-stratified treatment recommendations - an environmental scan of clinical practice guidelines. Bmc Medicine. 2013;11.
10. Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes. 2014;7(1):163-9.
11. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients - The need for risk stratification. Jama-J Am Med Assoc. 2007;298(10):1209-12.
12. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol. 2016;45(6):2075-88.
13. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 2010;11:85.
14. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages (vol 82, pg 661, 2004). Milbank Quarterly. 2006;84(4):759-60.
15. Kent S, Burn E, Dawoud D, Jonsson P, Ostby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2021;39(3):275-85.
16. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.
17. Rieke N, Hancox J, Li WQ, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. Npj Digital Medicine. 2020;3(1).
18. Dietterich TG. Ensemble methods in machine learning. Multiple Classifier Systems. 2000;1857:1-15.
19. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
20. 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(3):279-89.

21. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172(8):971-80.
22. Wolpert DH. Stacked Generalization. Neural Networks. 1992;5(2):241-59.
23. Williams RD, Markus AF, Yang C, Salles TD, DuVall SL, Falconer T, et al. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. medRxiv. 2020:2020.05.26.20112649.
24. Williams RD, Reps JM, Group OEKA, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. Knee Surg Sports Traumatol Arthrosc. 2022;30(9):3068-75.
25. John LH, Kors JA, Reps JM, Ryan PB, Rijnbeek PR. Logistic regression models for patient-level prediction based on massive observational data: Do we need all data? International Journal of Medical Informatics. 2022;163.
26. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56(9):826-32.
27. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.
28. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
29. van den Ham HA, Klungel OH, Singer DE, Leufkens HGM, van Staa TP. Comparative Performance of ATRIA, CHADS(2), and CHA(2) DS(2)-VASc Risk Scores Predicting Stroke in Patients With Atrial Fibrillation Results From a National Primary Care Database. Journal of the American College of Cardiology. 2015;66(17):1851-9.
30. Singer DE, Chang YC, Borowsky LH, Fang MC, Pomernacki NK, Udaltsova N, et al. A New Risk Scheme to Predict Ischemic Stroke and Other Thromboembolism in Atrial Fibrillation: The ATRIA Study Stroke Risk Score. Journal of the American Heart Association. 2013;2(3).
31. Wang TJ, Massaro JM, Levy D, Vasan RS, Wolf PA, D'Agostino RB, et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community - The Framingham Heart Study. Jama-J Am Med Assoc. 2003;290(8):1049-56.
32. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
33. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008;50(4):457-79.
34. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns. 2020;1(8).
35. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship (vol 15, 160018, 2016). Scientific Data. 2019;6.
36. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-75.
37. Registry NJ. National Joint Registry: National Joint Registry for England, Wales and Northern Ireland; 15th Annual Report. 2018.
38. Baker PN, Rushton S, Jameson SS, Reed M, Gregg P, Deehan DJ. Patient satisfaction with total knee replacement cannot be predicted from pre-operative variables alone A COHORT STUDY FROM THE NATIONAL JOINT

REGISTRY FOR ENGLAND AND WALES. Bone Joint J. 2013;95b(10):1359-65.

39. Maradit Kremers H, Larson DR, Crowson CS, Kremers WK, Washington RE, Steiner CA, et al. Prevalence of Total Hip and Knee Replacement in the United States. J Bone Joint Surg Am. 2015;97(17):1386-97.

40. Springer BD, Cahue S, Etkin CD, Lewallen DG, McGrory BJ. Infection burden in total hip and knee arthroplasties: an international registry-based perspective. Arthroplast Today. 2017;3(2):137-40.

41. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, et al. Mortality after surgery in Europe: a 7 day cohort study. Lancet. 2012;380(9847):1059-65.

42. Berstock JR, Beswick AD, Lopez-Lopez JA, Whitehouse MR, Blom AW. Mortality After Total Knee Arthroplasty: A Systematic Review of Incidence, Temporal Trends, and Risk Factors. J Bone Joint Surg Am. 2018;100(12):1064-70.

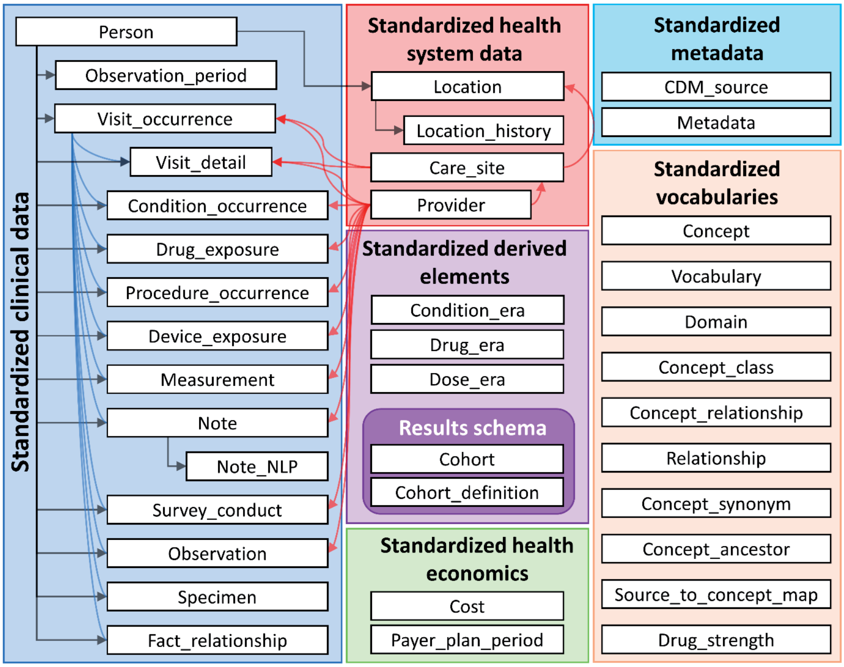
43. Listing J, Gerhold K, Zink A. The risk of infections associated with rheumatoid arthritis, with its comorbidity and treatment. Rheumatology (Oxford). 2013;52(1):53-61.

44. Dougados M, Soubrier M, Antunez A, Balint P, Balsa A, Buch MH, et al. Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA). Annals of the Rheumatic Diseases. 2014;73(1):62-8.

SAMENVATTING

In het hart van de geneeskunde ligt de besluitvorming. Elke dag staan clinici voor de complexe uitdaging om beslissingen te nemen over de behandeling van een patiënt. Als een patiënt bijvoorbeeld hoge bloeddruk heeft, welke behandeling moet hen dan kiezen (1)? Bij de besluitvorming over de behandeling van een patiënt worden de voordelen om een optimaal behandeltraject te kiezen vergeleken met de risico's van een behandeling (2). Wanneer een patiënt al een hoog risico op hartfalen loopt, zal het toevoegen van een geneesmiddel waarvan bekend is dat deze het risico op hartfalen verhoogt, waarschijnlijk worden afgeraden (3-6). De wetenschap dat een behandeling het risico op een bepaalde uitkomst verhoogt of verlaagt, is niet het enige relevante element van risico. Voor iedere patiënt kan een toename van het risico op één resultaat veel belangrijker zijn dan de toename van het risico op een ander resultaat. Dit is de afweging tussen relatieve en absolute risicotoename (7). Een kleine toename van een groot absoluut risico is waarschijnlijk een grotere, zorgwekkender toename dan een grote toename van een klein risico. Vaak wordt daarom in richtlijnen aanbevolen dat behandelingen verschillen op basis van het risico op bepaalde uitkomsten en de verschillende bekende risicoprofielen van medicijnen (8, 9). Een voorbeeld hiervan is gegeven in hoofdstuk 4, waarin de desbetreffende richtlijnen de keuze van een diabetesmedicijn afhangt van het risico op hartfalen. Dezelfde richtlijnen vermeldden echter niet hoe het risico van een patiënt moet worden berekend. Bij de toepassing van de richtlijnen moet de arts beoordelen wat het risico op een bepaalde gebeurtenis is voor een patiënt. Om deze beoordeling te maken, kan hij gebruik maken van informatie waarover hij beschikt, zoals causale gevolgtrekkingen, ervaring en klinische intuïtie. Het gebruik van causale inferentie studies brengt hen slechts zover. Deze studies geven een verschil in absoluut risico, uitgedrukt als relatief risico over een hele populatie. Dit gemiddelde risico houdt geen rekening met de heterogeniteit op patiëntniveau. Bovendien is de ervaring van een clinicus beperkt, aangezien hij niet altijd genoeg patiënten ziet om zeldzame patronen te herkennen. Bovendien wordt intuïtie beperkt door dingen die hij niet weet, zoals bijvoorbeeld de volledige voorgeschiedenis van de patiënt. Als zodanig is de in de richtlijnen aanbevolen risicobeoordeling op patiëntniveau een zeer moeilijke taak. Dit beperkt in ernstige mate de mogelijkheid om de behandeling van een patiënt te personaliseren. In het algemeen wordt de heterogeniteit op patiëntniveau als problematisch voor de behandeling beschouwd (10-14). Diezelfde heterogeniteit kan echter worden gebruikt voor personalisatie, als we die kunnen benutten in voorspellingsmodellen. Door een goed presterend voorspellingsmodel aan te bieden om de waarschijnlijkheid van een uitkomst te beoordelen, kunnen beter geïnformeerde beslissingen in de gezondheidszorg worden genomen en kan uiteindelijk het medische traject van een patiënt beter worden gepersonaliseerd. Deze personalisering zal bijdragen aan een beter presterend, efficiënter gezondheidszorgsysteem voor zowel de clinicus als de patiënt.

Om betere risicomodellen te maken, moeten we beschikken over gestandaardiseerde gegevens, gestandaardiseerde analytische verwekingstappen en gestandaardiseerde onder-



Figuur 1 De structuur van het OMOP-CDM

zoekspraktijken (15). Dit maakt een directere vergelijking van modellen mogelijk en zorgt voor een algemeen groter vertrouwen in de robuustheid van de gebruikte methodologie. De beste manier om deze standaardisaties door te voeren is door te beginnen met de gegevens. Verdere standaardiseringen van analytische verwerkingstappen wordt gemakkelijker wanneer de gegevens die we verzamelen voor observationeel onderzoek in kaart kunnen worden gebracht om de semantische en syntactische interoperabiliteit te verbeteren. Eén methode is het gebruik van het Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) dat wordt onderhouden door de Observational Health Data Sciences and Informatics (OHDSI) gemeenschap (zie Figuur 1) (16).

Het OMOP-CDM heeft als doel om zowel de syntactische als de semantische interoperabiliteit van observationele gezondheidsgegevens te verbeteren. Het standaardiseren van klinische gegevens naar een gemeenschappelijk formaat, weergegeven in Figuur 1 (blauw kader), maakt het gebruik van gestandaardiseerde analytische gereedschap mogelijk. Het gebruik van gestandaardiseerde vocabulaires (oranje kader) verbetert de semantische interoperabiliteit, dat wil zeggen het vergemakkelijkt de identificatie van klinische concepten met behulp van een gemeenschappelijke terminologie. Meer informatie over het OMOP-CDM staat in het boek van OHDSI (https://book.ohdsi.org).

Dit proefschrift beschrijft het werk dat de best practices helpt te definiëren voor het uitvoeren van Patient-Level Prediction (PLP) onderzoek met behulp van gegevens uit een federaal netwerk van gestandaardiseerde observationele databases.

Deel I

In deel I hebben wij bekeken hoe PLP modellen binnen gefedereerde gegevensnetwerken het best kunnen worden ontwikkeld en geëvalueerd. Een gefedereerd netwerk is een netwerk van databases die patiënten informatie beschermen door gebruik te maken van een op privacy gebaseerde aanpak (17). Aangezien veel gezondheidsgegevens in datasilos zijn ondergebracht, en niet buiten een specifieke omgeving mogen worden verspreid, is bundeling van gegevens onmogelijk. Een gefedereerd netwerk neemt veel van de problemen met de silo's van gegevens weg door in het hele netwerk een gemeenschappelijk gegevensmodel te implementeren om ervoor te zorgen dat de databases in hoge mate interoperabel zijn. Omdat we de onderliggende syntactische en semantische databasestructuur kennen, kunnen we standaard analyse-instrumenten delen tussen databases in plaats van de database door te geven aan het instrument. Dit betekent dat alleen software en samenvattende statistieken worden gedeeld en patiëntgegevens dus niet. Alle studies in dit proefschrift zijn uitgevoerd met gegevens die zijn gekoppeld aan de OMOP CDM. Zonder deze mapping zou dit werk vrijwel onmogelijk zijn geweest.

In hoofdstuk 1 is beoordeeld of de transporteerbaarheid van modellen kan worden verbeterd door verschillende modellen, elk ontwikkeld in een afzonderlijke database, te ensembleren en toe te passen in een nieuwe database (18, 19). Transporteerbaarheid wordt hier gedefinieerd als het vermogen van een model om zijn prestaties te handhaven wanneer het in een nieuwe database wordt beoordeeld (20, 21). Elke afzonderlijke database bevat vaak een populatie die niet representatief is voor de bredere ziekte specifieke populatie (de IBM Marketscan Medicare supplementaire database bevat bijvoorbeeld oudere patiënten die welvarender zijn). Wij onderzochten of wij modellen konden combineren die in verschillende databases zijn ontwikkeld (in plaats van soorten populaties) om een model te creëren dat beter is voor de algemene bevolking (en dus beter transporteerbaar is). Voor elke vraag in een set van 21 voorspellingsvragen trainden wij vijf basismodellen modellen uit één database, elk met behulp van een andere observationele gezondheidszorgdatabase. Vervolgens hebben wij verschillende ensemblemodellen ontwikkeld en onderzocht die de vijf verschillende modellen combineren, of *ensembleren*. Deze ensemblemodellen gebruikten de prestaties van de basismodellen (of niveau 1 modellen) en pasten vervolgens een van de verschillende fusiemethoden of een van de drie stapelmethoden toe. Bij fusiemethoden wordt de nieuwe patiënt gekoppeld aan het onderliggende model op basis van leeftijd. Bij een van de drie verschillende stapelmethoden wordt gebruik gemaakt van toenemende hoeveelheden gegevens. De stapelmethode past een logistisch regressiemodel toe op de outputs van de basismodellen om deze op een elegantere manier te combineren (22). Deze vorm van ensembling vereist gegevens van de validatieset. De prestaties van elk model zijn onderzocht via discriminatie en kalibratie met behulp van gegevens uit een nieuwe database die niet bij de modelontwikkeling werd gebruikt. De interne validatie van een model dat is ontwikkeld met behulp van de "hold out" database, werd berekend en ter vergelijking gepresenteerd als het "hypothetische optimum". Fusiemethoden presteerden over het algemeen beter dan de modellen met één database en waren consistenter bij toepassing op nieuwe gegevens. Stapelmethoden presteerden onvoldoende in termen van discriminatie wanneer de extra gegevens die nodig zijn voor het stapelen beperkt waren. De kalibratie was onvoldoende wanneer ensembles en modellen met één database werden toegepast op nieuwe databases. Bij vergelijking van alle hierboven beschreven methoden hebben wij geconstateerd dat ensemblemethoden over het algemeen beter presteren dan basismodellen. Deze prestaties moeten worden gezien in de context van een sterk toegenomen complexiteit van de modellen. Alle ensemblemodellen moeten opnieuw worden gekalibreerd voordat zij klinisch worden toegepast.

In hoofdstuk 2 werd het gebruik van proxy-leren en parsimonisatie overwogen (23, 24). Proxy-learning is de ontwikkeling van een model in een soortgelijke "proxy"-populatie als de belangrijkste doelpopulatie. Parsimonisatie is het proces waardoor modellen versimpeld worden. Dit model wordt vervolgens geëvalueerd in de echte probleemsetting in een nieuwe dataset. Dit wordt gedaan omdat er, om uiteenlopende redenen, onvoldoende gegevens kunnen zijn om in de werkelijke doelpopulatie te ontwikkelen (25). In hoofdstuk 2 gebeurde dit proces gevolgd, met het gebruik van gegevens van een andere ziekte. De standaardvorm van externe validatie is om deze validatie uit te voeren op een identieke probleemsetting, maar dan in een nieuwe database. In dit hoofdstuk wordt een model beschreven dat is ontwikkeld aan het begin van de pandemie van het Coronavirus (covid-19). De modellen moesten dan ook snel worden ontwikkeld op een moment dat er weinig gegevens over covid-19-infecties beschikbaar waren. Het doel was zo snel mogelijk een bruikbaar model te creëren. De parsimonisatie omvatte een proces in twee stappen waarbij eerst een data gedreven model werd ontwikkeld, vervolgens klinische expertise werd gebruikt om de covariaten te verfijnen tot een hanteerbaarder aantal en vervolgens de prestaties werden vergeleken. Bij het proxy-leren werd gebruik gemaakt van griepgegevens, die overvloedig aanwezig zijn, om te proberen snel een model te ontwikkelen dat voor covid-19-patiënten kan worden gebruikt. Wegens de beperkte hoeveelheid Covid-19-gegevens besloten wij deze proxy-methode te gebruiken om de Covid-gegevens te bewaren voor validatie. Dit zorgt om de bewijskracht van de prestaties te vergroten zonder langer te hoeven wachten om meer gegevens te verzamelen. Wij ontwikkelden drie modellen die ziekenhuisopname, ziekenhuisopname met intensieve diensten of overlijden en fataliteit beoordelen. Deze drie eindpunten vertegenwoordigen verschillende ziekte-ernstigheden. De drie modellen presteerden goed voor zowel complexe als versimpeld versies in de influenzadataset waarvoor we 150.000 patiënten bemonsterden. Deze prestaties werden grotendeels herhaald in de meerdere internationale Covid-19 datasets waarover wij beschikten (n=44.507). Deze analyse toonde aan dat proxy learning een efficiënte en effectieve techniek kan zijn. Door het gesimplificeerd karakter van de ontwikkelde modellen kunnen ze gemakkelijk in meerdere settingen worden gebruikt om de patiëntenzorg en de strategische planning te beïnvloeden.

Hoofdstuk 3 van dit proefschrift stelt een best practice vast voor het uitvoeren van externe validatie (26-28). Het hoofdstuk laat zien hoe de standaardisatie van gegevens naar het OMOP CDM helpt om de ontwikkeling van gestandaardiseerde instrumenten te vergemakkelijken. Verder, zorgen deze standaardisaties dat de last voor het extern valideren van een voorspellingsmodel in een netwerk van observationele databases is verminderd. Daartoe hebben wij meerdere modellen uit de literatuur genomen, deze geïmplementeerd in een OMOP CDM-compatibel formaat en toegepast in zes databases. Om het nut van deze verwerkingstappen aan te tonen, werd een use-case studie uitgevoerd met vijf bestaande modellen die incidentele beroerte bij patiënten met atriumfibrilleren voorspelden. De vijf bestaande modellen (ATRIA, $CHADS_2$, $CHADS_2$ VASC, Q-Stroke en Framingham) konden worden geïntegreerd in het OHDSI-kader voor voorspelling op patiëntniveau. Zo genoemde patient-level prediction. (29-31). Zij verkregen gemiddelde c-statistieken tussen 0,57-0,63 voor de zes databases. Dit was vergelijkbaar met andere validatiestudies. De validatienetwerkstudie werd binnen 60 dagen uitgevoerd over zes datasets nadat de modellen waren gerepliceerd. De technieken in dit hoofdstuk bleken een effectieve manier om modellen extern te valideren. De snelheid waarmee de modellen binnen het netwerk konden worden gevalideerd, nadat ze waren omgezet in een OMOP CDM-compatibel formaat, toont de kracht aan van de standaardisatie van analytische verwerkingstappen om snel en betrouwbaar bewijsmateriaal te genereren.

In hoofdstuk 4 hebben we bekeken hoe we de prestaties van een model het best kunnen beoordelen in een externe validatiesetting (32, 33). Dit hoofdstuk introduceerde het idee van prestatie in context. Om zinvolle uitspraken te kunnen doen over hoe de prestaties verschillen tussen interne en externe validatie, is het nuttig een verwachting te hebben van de prestaties in een nieuwe populatie (34). In dit hoofdstuk werd voorgesteld dit op twee manieren te doen. De eerste is het gebruik van een basismodel, wij stellen een eenvoudig leeftijds- en geslachtsmodel voor, dat context geeft aan de impact van toenemende complexiteit van het model. Als het basismodel vergelijkbare prestaties levert als het complexe model, dan weten we dat de extra complexiteit geen relevante prestatieverhoging oplevert. De tweede is het vaststellen van verwachtingen van een "volledig model" door een nieuw model te ontwikkelen in de externe valideringsdatabase om inzicht te geven in wat de prestatieverwachting in deze database is. Als een model een interne validatieprestatie van 0,7 heeft in database A, maar een externe validatie in database B van 0,6, ziet dit er onvoldoende uit. Als we echter weten dat een model dat in database B is ontwikkeld een AUC van 0,62 heeft, dan is de daling van de prestatie bij externe validatie waarschijnlijk te wijten aan case-mix. Met deze procedure hebben wij vastgesteld dat een database vaak moeilijker of gemakkelijker te voorspellen is. Deze variatie in moeilijkheidsgraad werd aangetoond aan de hand van een use-case van het voorspellen van incidenteel hartfalen bij diabetespatiënten in het eerste jaar na aanvang van een tweede diabetesmedicijn. In totaal werden 403.187 patiënten uit 5 databases in de studie opgenomen. Wij ontwikkelden 5 modellen die bij interne beoordeling een discriminerende prestatie hadden, beoordeeld aan de hand van de c-statistiek, variërend van 0,73 tot 0,81 en een aanvaardbare kalibratie. Bij externe validering van deze modellen in een nieuwe database presteerden drie modellen consistent en in de context vaak vergelijkbaar met modellen die in de database zelf waren ontwikkeld. Deze studie gaf niet alleen inzicht in de potentiële prestaties van het klinische model, maar ook in de databases zelf, waaruit bleek dat sommige moeilijker te voorspellen waren dan andere. Het proces waarbij ontwikkelings- en validatiedatabases worden gerouleerd en een basismodel wordt toegepast, wordt iteratieve paarsgewijze externe validatie (IPEV) genoemd. IPEV toont de potentiële toegevoegde waarde aan van het gebruik van complexere modellen en geeft context aan de modelprestaties in nieuwe databases.

In hoofdstuk 5 werd de ontwikkeling en invoering van de DELPHI-bibliotheek van voorspellingsmodellen gedetailleerd beschreven. Het hoofddoel hiervan was de verspreiding en evaluatie van voorspellingsmodellen die ontwikkeld zijn voor OMOP CDM-gegevens te vergemakkelijken (35). Gezien de snelle toename van de ontwikkeling van voorspellingsmodellen in de afgelopen tien jaar, maar het gebrek aan verbetering in de rapportage van modellen, was er een duidelijke onbeantwoorde behoefte aan een gecentraliseerde opslagplaats om het verspreidingsproces van modellen te standaardiseren. De wijdverspreide toepassing van de OMOP CDM en de gestandaardiseerde analyses, die mogelijk zijn gemaakt met het R-package PatientLevelPrediction (36), betekenen dat veel geproduceerde voorspellingsmodellen nu een gemeenschappelijk resultatenobject hebben. Door gebruik te maken van dit standaard resultatenobject kunnen modellen en hun prestaties gemakkelijk worden geformatteerd in een database. Wij ontwikkelden een database backend die de informatie over een model of de externe validatie van een model opneemt en opslaat. Bovendien hebben we een grafische gebruikersinterface (GUI) ontwikkeld waarmee gebruikers op een eenvoudige en intuïtieve manier met de database kunnen interageren. Belangrijk is dat hiermee een dynamische resultatenverkenningsomgeving wordt geladen. Met deze dynamische omgeving kunnen gebruikers van de GUI de resultaten verkennen en parameters wijzigen (bijvoorbeeld de drempel voor een beslissing) om te zien hoe dit de prestaties van het model beïnvloedt. De modellen zelf kunnen ook worden gedownload, geëvalueerd op nieuwe gegevens, en deze prestaties kunnen opnieuw worden geladen als een externe validatie. Het doel is dat externe validatie hierdoor gemakkelijker wordt en dus vaker wordt uitgevoerd. Op deze manier wil de DELPHI-bibliotheek de resultaten veranderen van een statisch object, bijvoorbeeld een tijdschriftartikel, in een dynamisch ecosysteem van bewijsvoering. Dit zou de schaal waarop voorspellingsmodellen worden gevalideerd moeten vergroten. De DELPHI-bibliotheek betekent een belangrijke stap in de implementatie van modellen in de klinische praktijk door de flexibiliteit van de evaluatie door clinici, toezichthouders en onderzoekers te verbeteren. Het is te hopen dat dit leidt tot meer vertrouwen op het gebied van PLP-modellering.

Deel II

In het tweede deel van dit proefschrift wordt de ontwikkeling van methoden achterwege gelaten en worden specifieke klinische vragen behandeld die werden beantwoord met een multidisciplinair team. De vragen kwamen rechtstreeks van clinici die een duidelijke behoefte zagen aan een voorspellingsmodel in hun dagelijkse praktijk.

Hoofdstuk 6 beschrijft de ontwikkeling en validatie van een voorspellingsmodel voor 90-dagen postoperatieve resultaten na een totale knieprothese (TKR). TKR is een veilige en kosteneffectieve chirurgische procedure voor de behandeling van ernstige knieartrose (OA) (37-39). Hoewel complicaties na een operatie zeldzaam zijn, zouden voorspellingsmodellen kunnen helpen bij het identificeren van patiënten met een hoog risico, voor wie preventieve maatregelen zouden kunnen worden genomen (40-42). Het doel was een eenvoudig model te ontwikkelen en te valideren dat kan helpen bij het maken van behandelkeuzes. Dit hoofdstuk bespreekt de ontwikkeling van een voorspellingsmodel voor mortaliteit dat werd uitgevoerd tijdens een study-a-thon in Oxford die meerdere artikelen opleverde. De belangrijkste bevinding van dit hoofdstuk was de ontwikkeling van een voorspellingsmodel voor sterfte binnen 90 dagen. Voor deze probleemstelling werden zowel een complex als een eenvoudig model ontwikkeld. Het complexe model had een c-statistiek van 0,78 intern en 0,70 extern. Het eenvoudige model had een c-statistiek van 0,77 intern en 0,71 extern. Hieruit blijkt dat het versimpeld model even goed presteert als het complexe model en als zodanig de voorkeur geniet. De prestaties van het versimpeld model wijzen erop dat het een klinische impact zou kunnen hebben. Het gebruik van slechts 12 variabelen in het model betekent dat het implementeerbaar is en onmiddellijk kan helpen bij de besluitvorming rond een operatie.

Voor hoofdstuk 7 hebben wij een voorspellingsmodel ontwikkeld voor patiënten met reumatoïde artritis (RA). De identificatie van RA-patiënten met een verhoogd risico op een of meer ongunstige gezondheidsuitkomsten blijft een grote uitdaging (43, 44). Dit hoofdstuk bespreekt de ontwikkeling en validatie van voorspellingsmodellen voor verschillende ongunstige gezondheidsuitkomsten bij RA-patiënten die beginnen met eerstelijns monotherapie van methotrexaat (MTX). De modellen zijn ontwikkeld en intern gevalideerd op 21.547 RA-patiënten en extern gevalideerd op 131.928 RA-patiënten. Modellen voor ernstige infectie (AUC: intern 0,74, extern variërend van 0,62 tot 0,83), MI (AUC: intern 0,76, extern variërend van 0,56 tot 0,82), en beroerte (AUC: intern 0,77, extern variërend van 0,63 tot 0,95), vertoonden voldoende discriminatie en adequate kalibratie. Modellen voor de andere uitkomsten vertoonden interne discriminatie met AUC < 0,65 en werden niet extern gevalideerd. Wij ontwikkelden en valideerden voorspellingsmodellen voor diverse ongunstige gezondheidsuitkomsten bij RA-patiënten die begonnen met eerstelijns MTX-monotherapie. De uiteindelijke modellen voor ernstige infectie, MI en beroerte bleken goed te presteren in meerdere databases. Deze modellen kunnen mogelijk helpen bij het personaliseren van de behandeling, bijvoorbeeld door profylactisch gebruik van antibiotica, meer toezicht op patiënten met een hoog risico en geruststelling van patiënten met een laag risico. Deze modellen zijn kandidaat om te worden bestudeerd voor klinisch gebruik.

Geconcludeerd kan worden dat voorspellingsmodellen van grote invloed kunnen zijn op de besluitvorming in de klinische praktijk en de personalisatie van de gezondheidszorg. Er blijven echter meerdere uitdagingen bij de ontwikkeling van effectieve en implementeerbare modellen. Om de complexe implementatietaak te rechtvaardigen, moeten de modellen een sterk vermogen aantonen om de patiëntenzorg positief te beïnvloeden. Een positieve invloed hebben op de patiëntenzorg moet het hoofddoel blijven van onderzoek naar voorspellingsmodellen op patiëntniveau.

BIBLIOGRAPHY

1. Kjeldsen S, Feldman RD, Liu LS, Mourad JJ, Chiang CE, Zhang WZ, et al. Updated National and International Hypertension Guidelines: A Review of Current Recommendations. Drugs. 2014;74(17):2033-51.
2. Muhlbacher AC, Juhnke C, Beyer AR, Garner S. Patient-Focused Benefit-Risk Analysis to Inform Regulatory Decisions: The European Union Perspective. Value in Health. 2016;19(6):734-40.
3. Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. Nat Rev Cardiol. 2011;8(1):30-41.
4. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, et al. 2022 AHA/ACC/HFSA Guideline for the Management of Heart Failure. Journal of Cardiac Failure. 2022;28(5):E1-E167.
5. Lippi G, Sanchis-Gomar F. Global epidemiology and future trends of heart failure. AME Medical Journal. 2020;5.
6. McDonagh TA, Metra M, Adamo M, Gardner RS, Baumbach A, Bohm M, et al. 2021 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure. European Heart Journal. 2021;42(36):3599-726.
7. Soares AA, Peto R. The big causes of death from noncommunicable disease. B World Health Organ. 2016;94(6):413-4.
8. Care F. Standards of Medical Care in Diabetes 2019. Diabetes Care. 2019;42(Suppl 1):S124-S38.
9. Yu T, Vollenweider D, Varadhan R, Li TJ, Boyd C, Puhan MA. Support of personalized medicine through risk-stratified treatment recommendations - an environmental scan of clinical practice guidelines. Bmc Medicine. 2013;11.
10. Burke JF, Hayward RA, Nelson JP, Kent DM. Using internally developed risk models to assess heterogeneity in treatment effects in clinical trials. Circ Cardiovasc Qual Outcomes. 2014;7(1):163-9.
11. Kent DM, Hayward RA. Limitations of applying summary results of clinical trials to individual patients - The need for risk stratification. Jama-J Am Med Assoc. 2007;298(10):1209-12.
12. Kent DM, Nelson J, Dahabreh IJ, Rothwell PM, Altman DG, Hayward RA. Risk and treatment effect heterogeneity: re-analysis of individual participant data from 32 large clinical trials. Int J Epidemiol. 2016;45(6):2075-88.
13. Kent DM, Rothwell PM, Ioannidis JP, Altman DG, Hayward RA. Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal. Trials. 2010;11:85.
14. Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages (vol 82, pg 661, 2004). Milbank Quarterly. 2006;84(4):759-60.
15. Kent S, Burn E, Dawoud D, Jonsson P, Ostby JT, Hughes N, et al. Common Problems, Common Data Model Solutions: Evidence Generation for Health Technology Assessment. Pharmacoeconomics. 2021;39(3):275-85.
16. Hripcsak G, Duke JD, Shah NH, Reich CG, Huser V, Schuemie MJ, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. Stud Health Technol Inform. 2015;216:574-8.
17. Rieke N, Hancox J, Li WQ, Milletari F, Roth HR, Albarqouni S, et al. The future of digital health with federated learning. Npj Digital Medicine. 2020;3(1).
18. Dietterich TG. Ensemble methods in machine learning. Multiple Classifier Systems. 2000;1857:1-15.
19. Reps JM, Williams RD, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.
20. 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(3):279-89.

21. Vergouwe Y, Moons KG, Steyerberg EW. External validity of risk models: Use of benchmark values to disentangle a case-mix effect from incorrect coefficients. Am J Epidemiol. 2010;172(8):971-80.
22. Wolpert DH. Stacked Generalization. Neural Networks. 1992;5(2):241-59.
23. Williams RD, Markus AF, Yang C, Salles TD, DuVall SL, Falconer T, et al. Seek COVER: Development and validation of a personalized risk calculator for COVID-19 outcomes in an international network. medRxiv. 2020:2020.05.26.20112649.
24. Williams RD, Reps JM, Group OEKA, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. Knee Surg Sports Traumatol Arthrosc. 2022;30(9):3068-75.
25. John LH, Kors JA, Reps JM, Ryan PB, Rijnbeek PR. Logistic regression models for patient-level prediction based on massive observational data: Do we need all data? International Journal of Medical Informatics. 2022;163.
26. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. J Clin Epidemiol. 2003;56(9):826-32.
27. Collins GS, de Groot JA, Dutton S, Omar O, Shanyinde M, Tajar A, et al. External validation of multivariable prediction models: a systematic review of methodological conduct and reporting. BMC Med Res Methodol. 2014;14:40.
28. Moons KG, Kengne AP, Grobbee DE, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: II. External validation, model updating, and impact assessment. Heart. 2012;98(9):691-8.
29. van den Ham HA, Klungel OH, Singer DE, Leufkens HGM, van Staa TP. Comparative Performance of ATRIA, CHADS(2), and CHA(2) DS(2)-VASc Risk Scores Predicting Stroke in Patients With Atrial Fibrillation Results From a National Primary Care Database. Journal of the American College of Cardiology. 2015;66(17):1851-9.
30. Singer DE, Chang YC, Borowsky LH, Fang MC, Pomernacki NK, Udaltsova N, et al. A New Risk Scheme to Predict Ischemic Stroke and Other Thromboembolism in Atrial Fibrillation: The ATRIA Study Stroke Risk Score. Journal of the American Heart Association. 2013;2(3).
31. Wang TJ, Massaro JM, Levy D, Vasan RS, Wolf PA, D'Agostino RB, et al. A risk score for predicting stroke or death in individuals with new-onset atrial fibrillation in the community - The Framingham Heart Study. Jama-J Am Med Assoc. 2003;290(8):1049-56.
32. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. Epidemiology. 2010;21(1):128-38.
33. Gerds TA, Cai T, Schumacher M. The performance of risk prediction models. Biom J. 2008;50(4):457-79.
34. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns. 2020;1(8).
35. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship (vol 15, 160018, 2016). Scientific Data. 2019;6.
36. Reps JM, Schuemie MJ, Suchard MA, Ryan PB, Rijnbeek PR. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. J Am Med Inform Assoc. 2018;25(8):969-75.
37. Registry NJ. National Joint Registry: National Joint Registry for England, Wales and Northern Ireland; 15th Annual Report. 2018.
38. Baker PN, Rushton S, Jameson SS, Reed M, Gregg P, Deehan DJ. Patient satisfaction with total knee replacement cannot be predicted from pre-operative variables alone A COHORT STUDY FROM THE NATIONAL JOINT

REGISTRY FOR ENGLAND AND WALES. Bone Joint J. 2013;95b(10):1359-65.

39. Maradit Kremers H, Larson DR, Crowson CS, Kremers WK, Washington RE, Steiner CA, et al. Prevalence of Total Hip and Knee Replacement in the United States. J Bone Joint Surg Am. 2015;97(17):1386-97.

40. Springer BD, Cahue S, Etkin CD, Lewallen DG, McGrory BJ. Infection burden in total hip and knee arthroplasties: an international registry-based perspective. Arthroplast Today. 2017;3(2):137-40.

41. Pearse RM, Moreno RP, Bauer P, Pelosi P, Metnitz P, Spies C, et al. Mortality after surgery in Europe: a 7 day cohort study. Lancet. 2012;380(9847):1059-65.

42. Berstock JR, Beswick AD, Lopez-Lopez JA, Whitehouse MR, Blom AW. Mortality After Total Knee Arthroplasty: A Systematic Review of Incidence, Temporal Trends, and Risk Factors. J Bone Joint Surg Am. 2018;100(12):1064-70.

43. Listing J, Gerhold K, Zink A. The risk of infections associated with rheumatoid arthritis, with its comorbidity and treatment. Rheumatology (Oxford). 2013;52(1):53-61.

44. Dougados M, Soubrier M, Antunez A, Balint P, Balsa A, Buch MH, et al. Prevalence of comorbidities in rheumatoid arthritis and evaluation of their monitoring: results of an international, cross-sectional study (COMORA). Annals of the Rheumatic Diseases. 2014;73(1):62-8.

DANKWOORD

Een jaar (te) laat, een taal geleerd en een nieuw thuis ontdekt en gemaakt. Na vijf jaar ligt mijn proefschrift klaar. There are many, many people I would like to thank for the support and friendship they showed me along the way.

Firstly, to my promotor **Peter**. The belief, faith and support you have shown me has been immeasurable. The trajectory of the PhD didn't always flow smoothly but you provided the space to try and fail and try again. This structured freedom gave me the room I needed to grow into the scientist I am today.

Jenna, you were added late as co-promotor but functioned as such from the first months. It is always inspiring to have extended discussions with you. The endless enthusiasm you have for pushing our research forward and your ability to start and finish projects off inspire me and I can but hope I gain some of that finishing ability in the next few years!

Ewout and **David**, thanks for the discussions and guidance, particularly in the early phases as we established the path for my PhD.

Tineke en **Desiree**, jullie zijn en waren fantastisch in het regelen van alles en nog wat op het afdeling. Werken met jullie maakt alles makkelijker. **Ilse**, jij bent ook een fijne aanvulling daarbij!

To my colleagues at Erasmus past and present, thanks for the collaboration and fun times. **Marten**, the first few months I remember how grateful I was that you took the time to explain to me how things worked (including but not limited to the human heart). **Esmé**, ik heb je altijd erg gewaardeerd als collega, en ik mis onze koffie "pauzes" die soms 2 uur lang waren. Dit is tijd waarin ik mijn mening over geneeskunde heb ontwikkeld, ondanks dat we eigenlijk nooit op een lijn eindigden… **Remy**, wat een eer was het om je paranymph te zijn! Je hebt me enorm geholpen gedurende het traject, ook wanneer je me hebt ondersteund toen ik besloot om terug naar het hotel te gaan om vers te kunnen presenteren in Philly… Een echt vriend door en door. **Emmely**, het is altijd gezellig om met jou te lunchen. Je hebt me veel geholpen en ondersteund tijdens promoveren. **Aniek, Cynthia** and **Tom**, jullie kwamen aan in het midden van mijn PhD en daar kwam covid nog steeds bij. Wat is het leuk om met jullie drieën te werken en eigenlijk nog leuker om met jullie te skiën; zelfs als we daarbij zeiknat worden. **Johan**, the department you had moulded was a great environment to learn in, where open discussion was encouraged and rewarded. Also, thanks for all the veg plants, I am eating the first of this years cucumber harvest as I write this! ! **Renske,** ik had nooit bedacht dat ik zou college geven zo leuk vinden. Dankjewel voor het vertrouwen en de aanmoediging om een eigen leerstijl te ontwikkelen.

Christel, Kiki, Natalie, Henrik, Alex, Marlies, Eliza, Katia, it was great to work with all of you and I hope to continue to run into you throughout our careers.

Patrick, thanks so much for the advice and the trust you show in me for the OHDSI community (also thanks for the data access). **Erica**, chatting with you is always a reason to look forward to an OHDSI event! And thanks to **Martijn, Clair** and **Sena** too. Without all the work done by you all and many more within OHDSI this PhD would've taken far far longer!

De bende van 8b+. **Bram, Laurens, Annelou** en **Zoë**, we gaan nooit 8b+ klimmen, maar we hebben tenminste mooie pofzakken. Jullie hebben mijn leven in NL verrijkt. Mooi klimmen, brandende zon of juist ijskoude regen, bloeden, rijden en skeer kamperen. Ik kijk ernaar uit om nog veel meer tripjes samen mee te maken en misschien zelfs een keer buiten te boulderen. **Pep** en **Sharon**, ook belangrijk klimmers, thanks voor de weekendjes weg.

Jongens van H13, **Niels**, **Chris** heel erg bedankt voor de lifts en de vriendschap. Leuk met jullie kampioenen te worden en op zondags te genieten van een potje hockey en een biertje.

Transacters, **Faizeh**, **Veronika**, **Pruthvi**, **Miguel**. Singapore, Bengaluru, Lausanne, Manchester, Rotterdam, Nijmegen, Lyon, Les Diablerets, Madrid, Leipzig and Berlin. You are missed, and meeting 6 monthly on works budget is also missed, but I couldn't have asked for a better start to working life than to have been added to a project with all of you. To see your struggles and then flourishes through your own PhD journeys was an inspiration as I stumbled through my own. Thank you guys for everything. And to the 3 honorary members **Priyanka**, **Guillaume**, and **Antoine**.

Eve, I realised writing this you're the person I have been friends with the longest. Thanks for taking the abstract descriptions I gave of networks and healthcare and turning it into the beautiful book this is.

The Lente boys, **Kris,** the true Rotterdam VIP, your reliability, and appetite for a party is something unmissably needed in my life. The memories we made with George and **Benny** in the wee hours are treasured if somewhat blurry. **George**, my paranymph, the truest and deepest of friendships. Its always a great comfort knowing that you're ready to jump on a train and come see me if the need ever arises. Thanks for the support over the years and your encouragement that I had made the right choices.

Katie, Alison, and **Ellen**, I never thought when we sat in first year mechanics that I would end up doing a PhD in machine learning, much less when we were struggling through Aristotle and Plato. Seeing how completely different the things we do now are always makes me smile.

Powell and **Olive**, I hope you both would have been proud of me.

Winifred, I never did work as hard as my sister but I did manage to get a PhD too.

Hannah, you're both brilliant and silly and I love you for that. Thanks for the lifelong support, academic and otherwise. You were invaluable over the last difficult phase of the PhD. **Matt**, like Hannah but sillier. Christmas in Paris with the both of you was unexpected but lovely and came in a difficult PhD moment and really helped me.

To my parents **Valerie** and **Steve**, I do not think I can express well enough my gratitude to you both. It wasn't always a smooth process with my academic career, but you both always supported me in the choices that I made and the encouragement you've always given me to take a jump and providing a safety net for my moves to both Brno and Rotterdam. This always gave me the feeling it was safe to try these things out. I love you both for that.

And finally, **Tiza**. I could have maybe got here without you maar het was wel veel minder leuk geweest. The person that gives me the most invaluable encouragement and support. You are always there to catch me (sometimes literally). The only person who can convince me to get me up at 4am. You share **Teddy** with me, who contributed absolutely nothing to the PhD and still he gets thanked. My partner in life and climbing, I love you.

CURRICULUM VITAE

Employment History

Doctoral Student

Erasmus University Medical Center [16/11/2017 – Present]
City: Rotterdam
Country: Netherlands

Marie Curie Early-Stage Researcher

Institute of Scientific Instruments [01/09/2015 – 30/09/2016]
City: Brno
Country: Czechia

Education and training

Data Science MSc

King's College, University of London [01/09/2016 – 31/01/2017]

Physics and Philosophy BSc

King's College, University of London [01/09/2012 – 01/08/2015]

PHD PORTFOLIO

Name:	Ross David Williams
Promotors:	Prof.dr.ir. Peter R. Rijnbeek Prof.dr. Ewout W. Steyerberg
Co-promotors:	Dr. Jenna M. Reps Dr.ir. David van Klaveren
Affiliation:	Erasmus University Medical Center
Department:	Medische Informatica

Description	**Organizer**	**EC**
Required		
Scientific Integrity (2018)	Erasmus MC Graduate School	1.00
2018 OHDSI Symposium (2018)	OHDSI	2.00
2019 OHDSI Europe Symposium (2019)	OHDSI	2.00
Health(y) Sciences Day (2019)	Erasmus MC	1.00
Oxford Real World Epidemiology summer School (2019)	NDORMS University of Oxford	2.00
35th ICPE Annual conference (2019)	INTERNATIONAL SOCIETY FOR PHARMACOEPIDEMIOLOGY	2.00
2019 OHDSI Symposium (2019)	OHDSI	2.00
OHDSI/EHDEN Study-athon (2020)	EHDEN	2.00
COVID-19 Study-A-Thon (2020)	OHDSI	2.00
2020 OHDSI Symposium (2020)	OHDSI	2.00
2022 OHDSI Symposium (2022)	OHDSI	2.00
2022 OHDSI Europe Symposium (2022)	OHDSI	2.00
OHDSI Community Calls (2023)	OHDSI	2.00
IPCI-Epi Meetings (2023)		2.00
Optional		
Supervised Student (2019)		5.00
Lecturer on Klinische Technologie (2021)	TU Delft	1.00
Lecturer on Klinische Technologie (2022)	TU Delft	1.00
Lecturer on Klinische Technologie (2022)	TU Delft	1.00
Supervised Student (2022)		5.00
Total EC		**39.00**

LIST OF PUBLICATIONS

Yang C, Kors JA, Ioannou S, John LH, Markus AF, Rekkas A, **Williams RD** et al. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. J Am Med Inform Assoc. 2022;29(5):983-9.

Williams RD, Reps JM, Kors JA, Ryan PB, Steyerberg E, Verhamme KM, et al. Using Iterative Pairwise External Validation to Contextualize Prediction Model Performance: A Use Case Predicting 1-Year Heart Failure Risk in Patients with Diabetes Across Five Data Sources. Drug Saf. 2022;45(5):563-70.

Williams RD, Reps JM, Group OEKA, Rijnbeek PR, Ryan PB, Prieto-Alhambra D. 90-Day all-cause mortality can be predicted following a total knee replacement: an international, network study to develop and validate a prediction model. Knee Surg Sports Traumatol Arthrosc. 2022;30(9):3068-75.

Williams RD, Markus AF, Yang C, Duarte-Salles T, DuVall SL, Falconer T, et al. Seek COVER: using a disease proxy to rapidly develop and validate a personalized risk calculator for COVID-19 outcomes in an international network. BMC Med Res Methodol. 2022;22(1):35.

Seinen TM, Fridgeirsson EA, Ioannou S, Jeannetot D, John LH, Kors JA, **Williams RD**, et al. Use of unstructured text in prognostic clinical prediction models: a systematic review. J Am Med Inform Assn. 2022.

Reps JM, **Williams RD**, Schuemie MJ, Ryan PB, Rijnbeek PR. Learning patient-level prediction models across multiple healthcare databases: evaluation of ensembles for increasing model transportability. BMC Med Inform Decis Mak. 2022;22(1):142.

Benz E, Wijnant SRA, Trajanoska K, Arinze JT, de Roos EW, de Ridder M, **Williams RD**, et al. Sarcopenia, systemic immune-inflammation index and all-cause mortality in middle-aged and older people with COPD and asthma: a population-based study. ERJ Open Res. 2022;8(1).

Reps JM, Kim C, **Williams RD**, Markus AF, Yang C, Duarte-Salles T, et al. Implementation of the COVID-19 Vulnerability Index Across an International Network of Health Care Data Sets: Collaborative External Validation Study. JMIR Med Inform. 2021;9(4):e21547.

Morales DR, Conover MM, You SC, Pratt N, Kostka K, Duarte-Salles T, **Williams RD,** et al. Renin-angiotensin system blockers and susceptibility to COVID-19: an international, open science, cohort analysis. Lancet Digit Health. 2021;3(2):e98-e114.

Burkard T, **Williams RD**, Vallejo-Yague E, Hugle T, Finckh A, Kyburz D, et al. Prediction of sustained biologic and targeted synthetic DMARD-free remission in rheumatoid arthritis patients. Rheumatol Adv Pract. 2021;5(3):rkab087.

Benz E, Trajanoska K, Schoufour JD, Lahousse L, de Roos EW, Terzikhan N, **Williams RD**, et al. Sarcopenia in older people with chronic airway diseases: the Rotterdam study. ERJ Open Res. 2021;7(1).

Wang Q, Reps JM, Kostka KF, Ryan PB, Zou Y, Voss EA, **Williams RD**, et al. Development and validation of a prognostic model predicting symptomatic hemorrhagic transformation in acute ischemic stroke at scale in the OHDSI network. PLoS One. 2020;15(1):e0226718.

Reps JM, **Williams RD**, You SC, Falconer T, Minty E, Callahan A, et al. Feasibility and evaluation of a large-scale external validation approach for patient-level prediction in an international data network: validation of models predicting stroke in female patients newly diagnosed with atrial fibrillation. BMC Med Res Methodol. 2020;20(1):102.

Burn E, Weaver J, Morales D, Prats-Uribe A, Delmestri A, Strauss VY, **Williams RD**, et al. Opioid use, post-operative complications, and implant survival after unicompartmental versus total knee replacement: a population-based network study. Lancet Rheumatol. 2019;1(4):E229-E36.

Benz E, Trajanoska K, Lahousse L, Schoufour JD, Terzikhan N, De Roos E, **Williams RD**, et al. Sarcopenia in COPD: a systematic review and meta-analysis. Eur Respir Rev. 2019;28(154).

ABOUT THE AUTHOR

Ross D. Williams was born on the 25th November, 1993 in London, UK. He grew up and attended secondary school in Manchester, graduating in 2012. He completed an undergraduate degree at King's College London in Physics and Philosophy (BSc) in 2015. After which he spent a year working as a Marie Curie scholar conducting research on NMR Spectroscopy in Brno, Czech Republic. He then returned to London to complete a master's in data science (MSc) again at King's, graduating in 2017. The same year, Ross started a PhD at Erasmus University Medical Center. For the past year, alongside his prediction research, Ross has been working on the DARWIN EU® project as analytics team lead and as lead of the personalised medicine work package in the EHDEN project. He co-leads the Patient-Level Prediction working group in OHDSI.

Despite living in the world's flattest country, Ross is a keen alpinist and rock climber who can often be found happily tied to a rope somewhere between south Belgium and north Italy.

Ross lives in Rotterdam with his partner Tiza and their cat Teddy.