Why predicting risk can't identify 'risk factors': empirical assessment of model stability in machine learning across observational health databases

Aniek F. Markus a.markus@erasmusmc.nl
Department of Medical Informatics
Erasmus University Medical Center
Rotterdam, The Netherlands

Peter R. Rijnbeek p.rijnbeek@erasmusmc.nl
Department of Medical Informatics
Erasmus University Medical Center
Rotterdam, The Netherlands

Jenna M. Reps jreps@its.jnj.com
Janssen Research and Development
Raritan, New Jersey, United States

Abstract

People often interpret clinical prediction models to detect 'risk factors', i.e. to identify variables associated to the outcome. We shed light on the stability of prediction models by performing a large-scale experiment developing over 450 prediction models using LASSO logistic regression and investigating model changes across databases (care settings) and phenotype definitions. Our results show that model stability, as measured by the similarity of selected variables, is poor across the prediction tasks but slightly better for the top (i.e. most important) variables. Differences in the top variables are mostly due to database choice and not due to using different target population and/or outcome phenotype definitions. However, this means using a different database might lead to finding different 'risk factors'. Furthermore, we found the effect (i.e. sign) of variables is not always the same across models, which makes clinical interpretation of potential 'risk factors' difficult. This study shows it is important to be careful when using LASSO regression to identify 'risk factors' and not to over-interpret the developed models in general. For 'risk factor' detection, we recommend investigating model robustness across settings or using alternative methods (e.g. univariate analysis).

1. Introduction

The implementation of supervised learning methods (e.g., logistic regression, gradient boosting machines, deep learning) on large observational healthcare data has lead to the development of prediction models that can calculate a patient's probability of experiencing future healthcare outcomes (see Yang et al. (2022) for a review of clinical prediction models). These prediction models have huge potential to improve clinical decision-making. There has also been interest in interpreting these prediction models to identify variables associated to the outcome, commonly referred to as 'risk factors'. Some researchers incorrectly interpret 'risk factors' as variables that cause the outcome, but it is a known misunderstanding that prediction models do not generally assess causality (Pearl, 2009; Schooling and Jones, 2018). However, even with a correct interpretation, the issue with using prediction models for 'risk factor' discovery is that prediction models are often developed using classification methods that optimize some objective function that measures similarity between the predictions and the ground truths. In other words, prediction models are developed with the aim to obtain the highest predictive performance rather than to identify a complete and consistent set of 'risk factors'.

Prediction models can leverage information available in routine-collected health care data, such as electronic health records or insurance claims. This data includes patient demographics and – depending on the database – the occurrence of medical conditions, drug prescriptions/dispensing, measurements, or procedures. As the data is typically sparse (binary variables indicate the presence of records in the medical file) and very high-dimensional ($>>$ 30,000 candidate variables), regularization is necessary to stop the model from overfitting by adding a cost to model complexity. One way of doing this is by selecting a subset of variables for the final prediction model. This is commonly done using the Least Absolute Shrinkage and Selection Operator (LASSO) regression, which reduces the size of the model by penalizing the sum of absolute coefficients.

LASSO regression can be thought of as performing variable selection during model fitting. Various forms of variable selection or regularization are commonly applied when researchers develop clinical prediction models using observational data. People often interpret the final model or set of variables selected into the model to identify an outcome's 'risk factors' (e.g. Nusinovici et al. (2022), see Section 2 for more examples). However, little is known about the stability of prediction models in practice (e.g., for LASSO logistic regression the final set of variables included in the model may differ based on subjective study design choices). If a model is unstable, using it for 'risk factor' detection is questionable. It is even more problematic if an unstable prediction model is interpreted to assess the effect of 'risk factors'.

Traditionally, stability in machine learning (ML) is defined as the robustness of the chosen variable set to differences in training data when drawn from the same generating distribution or population (Kalousis et al., 2005). Although relevant for the medical domain, we recognize differences in data might not only arise because of sources of instability such as noise, data dimensionality/sample size, imbalance of data, and variable redundancy (Nogueira et al., 2018). Study design choices such as the selected target population and outcome might be even more influential. A prediction model can be defined by specifying three components: 1) the target population (the patients you want to predict risk for at which point in time), 2) the outcome (what you are trying to predict), and 3) the time-at-risk (the time horizon within which the outcome should occur) (Reps et al., 2018). The first two components require phenotype definitions in observational data. There are often multiple ways to identify the target population/outcome and different researchers often use different phenotypes (e.g. Mentz et al. (2016)). In general, there is a trade-off between labelling patients with a condition correctly (positive predictive value) and detecting all patients with a condition (sensitivity). Phenotypes form the basis for any prediction model, however, the impact of differences in definitions on the resulting models is not well-studied.

We aim to perform a large-scale experiment developing over 450 prediction models using LASSO logistic regression to empirically evaluate the stability of prediction models trained using observational data. This will provide insight into the potential issues with interpreting prediction models for 'risk factor' discovery. We will develop prediction models using the Observational Health Data Sciences and Informatics (OHDSI) Patient-Level Prediction framework (Reps et al., 2018). The international research collaboration known as OHDSI has developed open-source data standards and tools that allow prediction models to be developed and externally validated rapidly, at a large scale, following accepted best practices (Khalid et al., 2021).

Generalizable Insights about Machine Learning in the Context of Healthcare

- It is known some algorithms are not stable, but there is very little attention to the impact of this 'problem' on developed clinical prediction models: "Is the variation in selected variables large or reasonably small? Is there a relation between the prediction task and resulting model instability? Would different 'risk factors' be identified?"
- In the current work, we shed light on the stability of prediction models in a clinically meaningful way by investigating the changes across databases (care settings) and phenotype definitions. We propose three intuitive steps to assess the stability of prediction models that are linear in the variables.
- We empirically show that a higher number of outcome cases leads to more variables being selected using LASSO logistic regression, but more stability in the variables (i.e. less different variables selected). The stability of the developed models is poor overall, but slightly better for the top (i.e. most important) variables. The database choice is important for the selected top variables; different databases lead to different 'risk factors'. Finally, interpreting the effect of 'risk factors' is problematic as the sign can differ across models.
- Researchers should be careful not to over-interpret prediction models as the identified 'risk factors' appear to depend on the study design choices. Therefore, we recommend investigating model robustness across settings or using other techniques for 'risk factor' detection (e.g. univariate analysis).

2. Related Work

Identifying Risk Factors

When researchers are interested in risk factors, they are generally interested in finding variables that are associated to the outcome. We define a 'risk factor' as a variable that is associated with the outcome. Historically, prediction models were developed by experts handpicking a small number (5-10) of variables, removing correlated variables, and then fitting a model. These models would then be interpreted to determine the 'risk factors' and their effects. Recently, an increase in the availability of large observational healthcare data has resulted in more advanced, data-driven, machine learning approaches to prediction model development. These models, developed with hundreds or thousands of variables, are often also being interpreted for 'risk factor' discovery. However, these methods generally do not account for correlations between variables, which makes interpretation problematic.

In healthcare data, we expect correlation between variables as procedures are often used to diagnose medical conditions, and being diagnosed with a medical condition often leads to a drug being prescribed. This means sets of procedures, medical conditions, and drugs are commonly observed together. If a data-driven approach to develop prediction models is applied, this correlation is likely to impact the observed effect of a variable.

Prediction models generally find variables that are associated with the outcome, however, the set of variables identified as predictive by a prediction model may only be a subset of the complete set of 'risk factors'. For example, LASSO regression often ignores variables that are weakly associated with the outcome. In addition, if two variables are highly correlated, LASSO regression is likely to ignore one of them or the correlation may cause interpretation issues with the coefficients. If a LASSO regression is interpreted to identify 'risk factors', then many real 'risk factors' may go undetected. An additional concern is that if models are unstable, then the set of identified 'risk factors' may vary based on how the model was developed, which is problematic, as you ideally want a consistent set of 'risk factors'. Even more problematic is interpreting the coefficients of a LASSO regression model (or variable importance of a ML model), as correlations between variables can cause these models be unstable. This makes it possible to find both positive and negative associations between a variable and an outcome, from which wrong conclusions may be drawn.

Despite these concerns, prediction models developed by algorithms such as LASSO regression or another form of regularized regression are commonly being interpreted for 'risk factor' discovery. For example, to identify 'risk factors' associated with (nine) mayor eye diseases (Nusinovici et al., 2022), COVID-19 mortality (Zhang et al., 2021), suicide attempts (García de la Garza et al., 2021), (self-reported) breast cancer (McEligot et al., 2020), change in postoperative pain outcomes (Parthipan et al., 2019), and inflammation in Crohn's disease patients (Reddy et al., 2019). Earlier stability analysis in the context of healthcare has investigated model stability and how to develop (more) stable models given a set of data (Gopakumar, 2017), in this work we investigate the stability when the data might change as a result of study design choices.

Stability Algorithms

It is known some algorithms that reduce the size of the model, such as LASSO regression, are not consistent variable selectors and adjustments might be needed (Leng et al., 2006). Different adjustments have been proposed in the ML literature such as choosing the tuning parameter differently (Meinshausen and Bühlmann, 2004), stability selection (Meinshausen and Bühlmann, 2010), or random lasso (Wang et al., 2011). Nevertheless, ordinary LASSO regression is commonly used (without adjustments) to develop prediction models for the medical domain (Christodoulou et al., 2019).

Measuring Stability

Stability of variable selection methods can be assessed by generating slightly different datasets (e.g. by taking bootstrap samples, noise injection, or random subsampling), then applying the variable selection method of interest and then measuring the variability in the selected variable set. For a review of measures to quantify stability we refer to Nogueira et al. (2018). Existing methods can be grouped in similarity-based versus frequency-based measures. Similarity-based measures define stability as the average pairwise similarity between all possible pairs of variable sets. These measures depend on the choice of similarity definition, e.g. the Hamming distance (Dunne et al., 2002), the Jaccard index (Kalousis et al., 2005), or the POG index (Shi et al., 2006). Frequency-based measures are (a function of) the observed frequencies of selection of each variable (or variable set). Some examples include Davis' Measure (Davis et al., 2006), Krizek's measure (Křížek et al., 2007), and Goh's measure (Goh and Wong, 2016). Based on their literature review, Nogueira et al. (2018) propose the 'stability estimator' as a novel (and preferred) measure to quantify stability.

3. Methods

3.1. Datasets

We used seven large observational healthcare databases in this study. All datasets used in this paper were mapped into a data structure known as the Observational Medical Outcomes Partnership Common Data Model (OMOP-CDM) (Overhage et al., 2012). The OMOP-CDM was developed to standardise local data in a consistent structure and vocabulary, which allows us to perform observational research at scale by utilising existing tools and sharing analysis code across sites. Database details are available in Table 1 and a list of descriptions in Appendix A. All databases obtained institutional review board approval when necessary.

3.2. Specification of Prediction Tasks

Following the proposed prediction framework by Reps et al. (2018), we defined a prediction question as: "Among a *target population*, which patients will develop an *outcome* during a *time-at-risk* period relative to index?". In this study we focused on developing models for nine different prediction tasks (corresponding to nine outcomes).

Target population
The target population of interest is the general population, we investigated the choice of three phenotype definitions:

T1. Patients with a healthcare visit during 2017 (index is a random visit in 2017)

T2. Patients with a influenza vaccination during 2017 (index is the date of the influenza vaccination in 2017)

T3. Patients observed in the database during 2017 (index is the 1st of January 2017)

Outcomes
The nine outcomes of interest are: acute myocardial infarction, anaphylaxis, appendicitis, disseminated intravascular coagulation, encephalomyelitis, Guillain-Barré syndrome, hemorrhagic stroke, non-hemorrhagic stroke, and pulmonary embolism. These outcomes were chosen as they were listed as COVID-19 vaccine outcomes of interest by the U.S. Food and Drug Administration (2021). For each outcome we used 3 or 4 different phenotype definitions (see Appendix B): a broad definition (including a wide range of diagnosis codes), a narrow definition (including a smaller set of diagnosis codes selected by the U.S. Food and

Table 1: List of databases included in study (ordered by database population size).

Database name	Acronym	Country	Data type	Population size	Age range
IBM MarketScan® Commercial Claims and Encounters Database	CCAE	USA	Claims	157M	0-65
Optum® De-identified Electronic Health Record Dataset	Optum EHR	USA	EHR	101M	All
Optum® De-Identified Clinformatics® Data Mart	Optum DoD	USA	Claims	91M	All
IBM MarketScan® Multi-State Medicaid Database	MDCD	USA	Claims	33M	All
IQVIA Disease Analyzer Germany	IQVIA Germany	Germany	Claims	31M	All
Japan Medical Data Center	JMDC	Japan	Claims	13M	All
IBM MarketScan® Medicare Supplemental Database	MDCR	USA	Claims	10M	Mostly 65+

Drug Administration), a narrow definition limiting outcomes to during inpatient and/or emergency department visits.

Time-at-risk
The time-at-risk period is one year for all prediction tasks as we were interested to investigate if we could predict outcomes associated with vaccines. We predicted the first time occurrence of the nine outcomes from 1 day until 365 days after the specified index date.

We chose to investigate multiple target population/outcome phenotype definitions across databases to see what impact, if any, these choices make on model stability, specifically the different model variables for the same prediction task. We selected these prediction tasks related to COVID-19 vaccination as models to predict negative outcomes were needed for the general population (as the vaccination was first recommended for all adults and later also for children), at a time when there were many unknowns and a lack of standards. This means the choice of target population and outcome definitions is less clear. Hence, this resembles a real-world setting where there was likely to be subjectivity in the study design choices when developing models.

3.3. Data Extraction

We created a labelled data set for each combination of database d, target population t, and outcome o:

- Databases $d \in D$ where D = {CCAE, Optum EHR, Optum DoD, MDCD, IQVIA Germany, JMDC, MDCR}.
- Target population $t \in \{T_1, T_2, T_3\}$, where T = {General population} with 3 different phenotype definitions.
- Outcome $o \in O$ where O = {Acute myocardial infarction, Anaphylaxis, Appendicitis, Disseminated intravascular coagulation, Encephalomyelitis, Guillain-Barré syndrome, Hemorrhagic stroke, Non-hemorrhagic stroke, Pulmonary embolism}. Each outcome o has 3 or 4 (p_o) different phenotype definitions $\{o_1, .. , o_{p_o}\}$.

Within each database d, we extracted a sample for each target population definition $\{T_1, T_2, T_3\}$. For each cohort, the following inclusion criteria were applied:

- Patients must be observed in the database >=365 days prior to index date

The index dates were a random visit during 2017 (T_1), the date of the influenza vaccination in 2017 (T_2), and 1st of January 2017 (T_3). As the cohorts were large we took a random 2 million patient sample for each database if the cohort was larger than 2 million.

For each target population cohort sample t in database d, we then extracted the following candidate variables:

- Age at index date in 5-year buckets (0-4, 5-9, 10-14, etc.)
- Sex
- One-shot encoding for any medical condition recorded in the 365 days up to 1 day prior to index date
- One-shot encoding for any drug recorded in the 365 days up to 1 day prior to index date

This resulted in a K-dimensional vector of candidate variables $\mathbf{x}_i^{dt}$ for patient i in database d and target population cohort sample t. Age and gender are required by the OMOP-CDM and were never missing. For conditions and drugs, no record was interpreted as patient does not have the condition or receive the drug (thus effectively imputed as zero). We used one-year covariate lookback as the choice of covariate lookback has been shown to have little impact on performance (Hardin and Reps, 2021) and using one-year covariate lookback has the advantage that variables are available for all patients.

For each outcome definition o_p we then extracted the outcome label for each patient in each database/target population by determining whether they had the outcome recorded during the time-at-risk (1 day after index until 365 days after index). If they did, they were assigned a label of 1 ($y_i^{dto_p} = 1$, indicating they had the outcome o_p during the time-at-risk) and if they did not, they were assigned a label of 0 ($y_i^{dto_p} = 0$).

This resulted in 588 (7 database, 3 target populations, and 28 outcome definitions) labelled data sets, where the data for database d, target population t and outcome definition o_p is $D^{dto_p} = \{(\mathbf{x}_i^{dt}, y_i^{dto_p})\}_{i=1}^N$.

3.4. Model Development

For each labelled data set D^{dto_p} per target population and outcome definition in each database, a random sample containing 75% of patients (‘training set) was used for development and the remaining 25% of patients (‘test set) was used for (internal) validation. Within each training set, we used 3-fold cross validation to pick the optimal regularization hyperparameter for LASSO logistic regression (Suchard et al., 2013). The hyperparameter selection optimised a ranking measure known as the area under the receiver operator curve (AUC). We then obtained the final prediction model by training the LASSO logistic regression with the optimal hyperparameter on the complete training set.

For each of the nine different prediction tasks, we had ∼50 models different models $\{f^{dto_p}\}$ that were learned using different combinations of database (d), target population (t) and outcome definition for the same outcome of interest (o_p). The developed models f^{dto_p} have the form:

$$\text{logit}(y_i^{dto_p}) = \Lambda^{-1}(y_i^{dto_p}) = \mathbf{x}_i^{dt}\hat{\beta}_i^{dto_p},$$

where $\Lambda(.)$ is the logistic distribution function and $\hat{\beta}_i^{dto_p}$ is a vector of coefficients $\{\hat{\beta}_{i1}^{dto_p}$, $\hat{\beta}_{i2}^{dt}$, ..., $\hat{\beta}_{iK}^{dto_p}\}$. Full code to develop the prediction models including target population/outcome definitions and data extraction code is available on GitHub: `https://github.com/ohdsi-studies/Covid19VaccinePrediction`.

3.5. Stability Metrics

As discussed in Section 2, different metrics have been proposed in the literature to quantify model stability. These metrics typically summarize model stability with a single number, which allows for a simple and quick comparison between different ML algorithms. However, these metrics are not always intuitive.

We propose three intuitive steps to assess model stability for models that are linear in the variables (like the developed models f^{dto_p} in our study):

1. *How many variables are selected across models?* We assessed differences in model size by calculating the number of non-zero coefficients per model ($\sum_{k=1}^{K} I[\hat{\beta}_{ik}^{dto_p} \neq 0]$ for all $d \in D$, $t \in T$).

2. *Are the same or different variables included across models?*

   a. *Full model* We assessed the stability of the chosen variable set using the stability estimator proposed by Nogueira et al. (2018) for all prediction tasks ($o \in O$):

   $$\Phi_o = 1 - \frac{\frac{1}{K}\sum_{k=1}^{K} s_k^2}{\frac{\bar{k}}{K}(1 - \frac{\bar{k}}{K})},$$

   where K is the total number of candidate variables, $\bar{k}$ is the average number of selected variables $\frac{1}{F}\sum_{f=1}^{F}\sum_{k=1}^{K} I[\hat{\beta}_k^f \neq 0]$ across F variable sets, and s_k^2 is the unbiased sample variance $\frac{F}{F-1}\hat{\rho}_k(1 - \hat{\rho}_k)$ of the selection of variable k. We opted for this metric as it can cope with any collection of variable sets (with varying number of variables), there is a known and finite range of values (0-1) which makes the interpretation meaningful, and it has a correction for chance (only reflecting systematic similarity in variable sets and not due to randomness) (Nogueira et al., 2018). Other metrics quantifying the stability of variable selection algorithms can be used if desired.

b. *Top 10/25 variables* Most literature focuses on assessing the stability of the full variable set, however, most attention is usually on the top (i.e. most important) variables when interpreting models. Hence, we analyzed the top 10/25 variables (as defined by largest absolute coefficients in LASSO) in more detail using the same metric introduced above.

   Besides that, we computed the overlap in the top 10 variables within each prediction task. We performed a pairwise comparison between models, counting the number of same variables among the most important 10 variables between each pair of models. Note this is equivalent to the (similarity-based) Percentage of Overlapping Genes (POG) index measure proposed by Shi et al. (2006), as the number of selected variables is constant across comparisons ($k = 10$). We compared differences between the same/different databases and phenotype definitions.

3. *Is the direction of the effect of variables the same across models?* Finally, differences in the model coefficients and in particular the direction of effect are important for model interpretation. A model coefficient represents the additive effect of a certain variable given all other variables. If one of the variables in a model is removed, all other model coefficients change and could even change direction. Therefore, we compared the sign of each variable across models. For each variable occurring in at least three models, we checked if the sign of the coefficient $\hat{\beta}_{ik}^{dto_p}$ was the same or different across models. We then computed the percentage of variables found with the same sign across models. We also studied the variation in signs for the 100 most often selected variables across models for each prediction task.

We used the above steps to evaluate the stability of models within each prediction task. Finally, we linked the stability to the internal and external discriminative performance of models as measured by the AUC.

4. Results on Real-world Data

We developed prediction models for 9 outcomes of interest x 3-4 phenotype definitions x 3 target population definitions x 7 databases to evaluate the stability of models. Stability was assessed in terms of the number of selected variables (Section 4.2), the similarity of selected variables (Section 4.3), and the sign of variables across models (Section 4.4). We end by investigating the clinical impact of model stability by evaluating the internal and external predictive performance of models.

4.1. Prediction Tasks Data Size

The size of the target cohort and corresponding number of outcome cases across the developed prediction models are summarised in Figure 1. The number of patients in the target populations was relatively large and stable (on average $1,678,441$-$1,824,745$ across prediction tasks), but the number of outcomes varied largely (on average 103-$7,790$ across prediction tasks).

As a result, not all databases were suitable to develop prediction models for each prediction task. For example, IQVIA Germany had insufficient numbers for some outcome definitions (20 out of 28) and JMDC did not contain information on influenza vaccination (thus misses 1 target population). A total of 457 models have been developed successfully, they can be explored online: `http://data.ohdsi.org/PatientLevelPredictionRepository`.

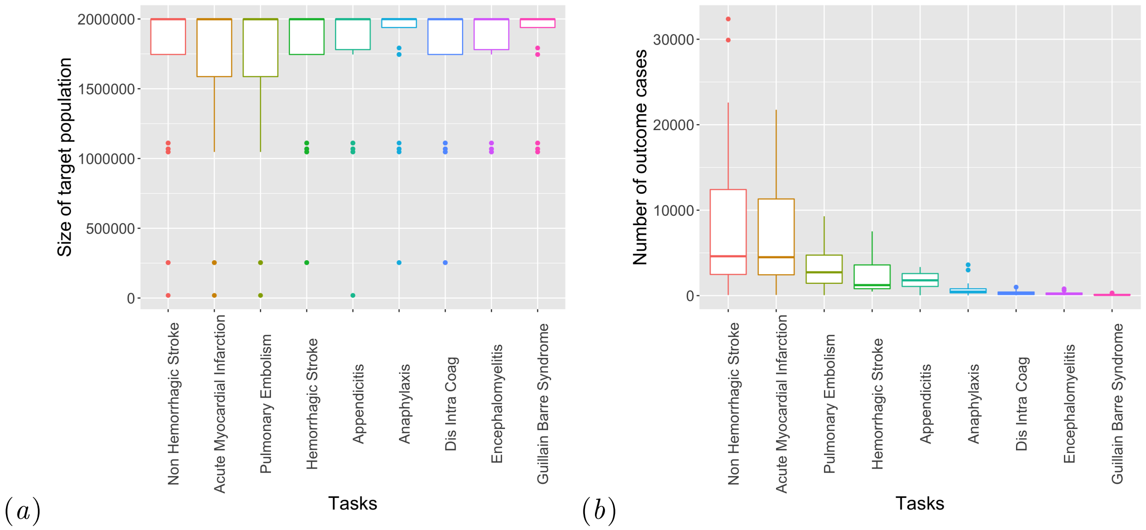

Figure 1: Boxplot of the size of the target population (a) and number of outcome cases (b) across databases and phenotype definitions showing large differences between prediction tasks, especially in the number of outcomes.

4.2. How many variables are selected across models?

The impact of using different databases and phenotype definitions on the size of models (as measured by the number of variables) are shown in Figure 2. There was a large variation in the number of selected variables, from a minimum of 2 variables (for anaphylaxis, appendicitis, disseminated intravascular coagulation, and pulmonary embolism) to a maximum of 1436 variables (for non-hemorrhagic stroke). Figure 2b shows a higher number of outcome cases generally leads to more variables being selected using LASSO logistic regression. However, comparing the number of variables against the size of the target population did not indicate a clear trend (see Figure 2a). We found that a smaller target population often has a smaller model, but not that a larger target population necessarily leads to a larger model (here model size varies and might be more related to the number of outcome cases).

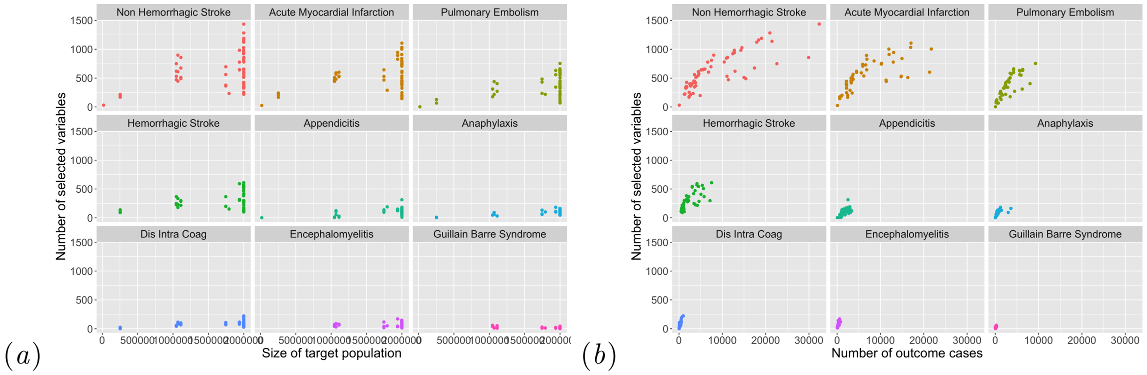

Figure 2: Scatterplot of the size of the target population (a) and number of outcomes (b) versus the number of selected variables per prediction task (ordered by a decreasing number of outcomes). Each point represents a developed model.

To further investigate the relation between the number of outcomes and model size, we split the results per database (see Figure 7 in Appendix C). This demonstrates that phenotype definitions including a higher number of outcomes result in larger models also within the same database.

4.3. Are the same or different variables included across models?

Next, we investigated the stability of the chosen variable set. Figure 3 visualizes model stability across prediction tasks using the stability estimator Φ (see Section 3.5). We observed overall model stability is relatively poor, with a maximum value of 0.18 across prediction tasks (maximum stability is 1). The top 10/25 variables, however, were slightly more stable with a maximum value of 0.44. This means there is less variation in the variables selected as top variables as compared to the entire model.

Again, we observed a relation with the number of outcomes. A higher average number of outcomes is positively correlated with stability of the chosen variable set (Pearson's ρ=0.89, 95% CI [0.54,0.98]). We further observed a larger average size of the target population is negatively correlated with stability of the chosen variable set (Pearson's ρ=-0.85, 95% CI [-0.97,-0.44]).

A more in-depth analysis of the top 10 variables is shown in Figure 4. Note that we did not investigate within-sample variance (same D, T, O) as further discussed in Section 5. This shows the top 10 variables for non-hemorrhagic stroke and the top 10 variables for acute myocardial infarction are quite similar across models developed within the same database, even when a different target population and/or outcome definition was used. However, this is not the case for the tasks predicting encephalomyelitis and Guillain-Barré syndrome, where we see hardly any overlap, even within the same database. For the prediction tasks in this study, this suggests that the impact of different target population and/or outcome definitions was limited when the number of outcome cases is sufficiently high. However, the influence of database choice was substantial, as can be seen from the low overlap across all

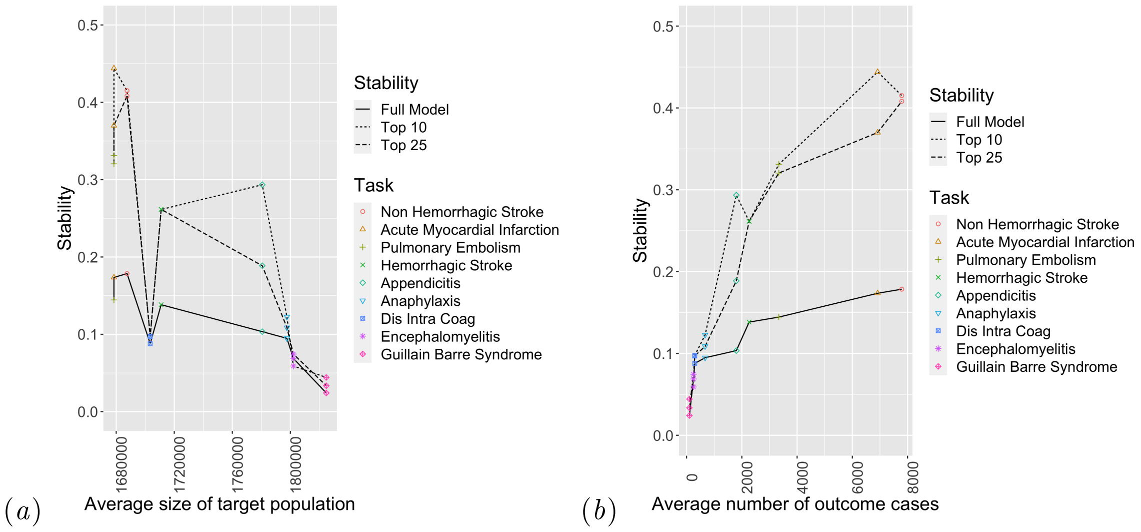

Figure 3: Graph visualizing model stability Φ as defined by Nogueira et al. (2018) versus the average size of the target population (a) and average number of outcome cases (b) across prediction tasks.

prediction tasks for models developed on different databases. This means that the top (i.e. most important) variables differed across databases.

A list of the 10 most often selected variables across models for each prediction task is included in Appendix C.

4.4. Is the direction of the effect of variables the same across models?

Finally, we investigated how often the sign of the coefficient for a given variable is the same (i.e. always positive or negative) across prediction tasks. We computed the percentage of variables with the same sign across prediction tasks (see Appendix C), we found this is highest for disseminated intravascular coagulation (54%) and lowest for non-hemorrhagic stroke (27%). The prediction tasks with a lower average number of outcomes, have a slightly higher percentage of same sign variables. This can be explained by the fact that these models were less stable, leading to more variation in the selected variables, and therefore less often contained the same variables (of which we could investigate the similarity of signs). Figure 5 shows the variation in signs for the 100 most often selected variables for each prediction task. Bars that are a single color indicate consistency in the sign of the coefficient (always a positive coefficient across models or always a negative coefficient across models). Bars that are green and red indicate that the coefficient was positive in some models and negative in others. Overall, it seems that the sign of the coefficient can vary greatly even for the top variables. We do not observe clear difference between prediction tasks, but less selected variables seem more likely to flip sign. These results clearly highlight the issue of interpreting the developed prediction models for ‘risk factor’ effect, as the effect often alternates between a positive and negative association.

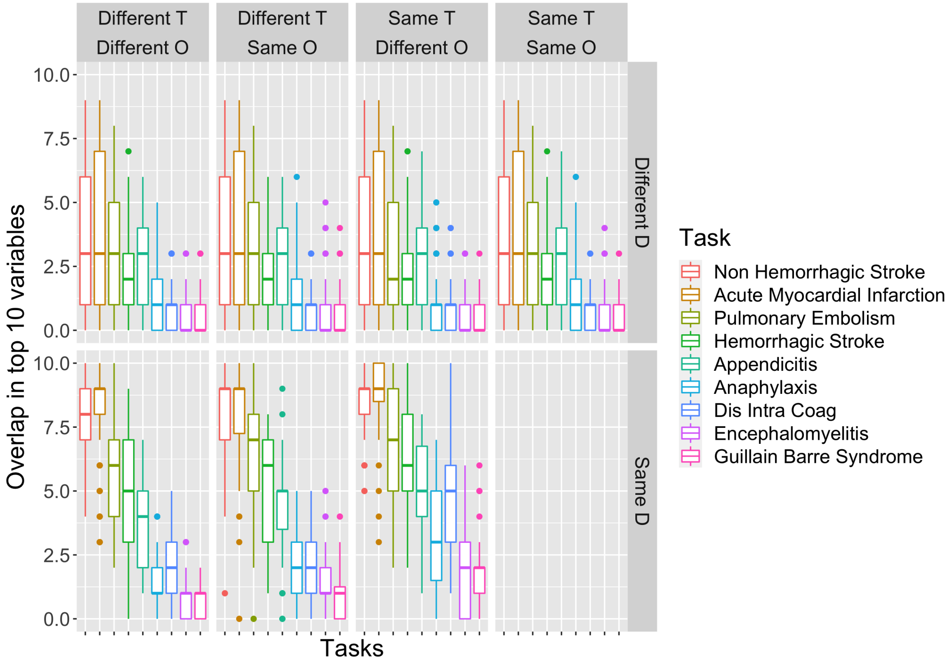

Figure 4: Boxplots showing the overlap in the top 10 variables as defined by counting the number of common variables between each pair of models for same/different database (D), target population definition (T), outcome definition (O) across models. Each point represents a developed model, colors indicate the corresponding prediction task (ordered by a decreasing number of outcomes).

4.5. Model Performance

Finally, we investigated the relation between stability and discriminative performance (AUC) for the nine prediction tasks (see Figure 6). The average AUC was slightly lower and more variable in new databases (external validation) compared to the same database (internal validation), but overall the performance was comparable. Prediction tasks for which the models were more stable had a higher internal (Pearson's ρ=0.76, 95% CI [0.21,0.95]) and external (Pearson's ρ=0.72, 95% CI [0.10,0.94]) predictive performance on average. However, both performance and stability may be influenced by the complexity of the prediction task.

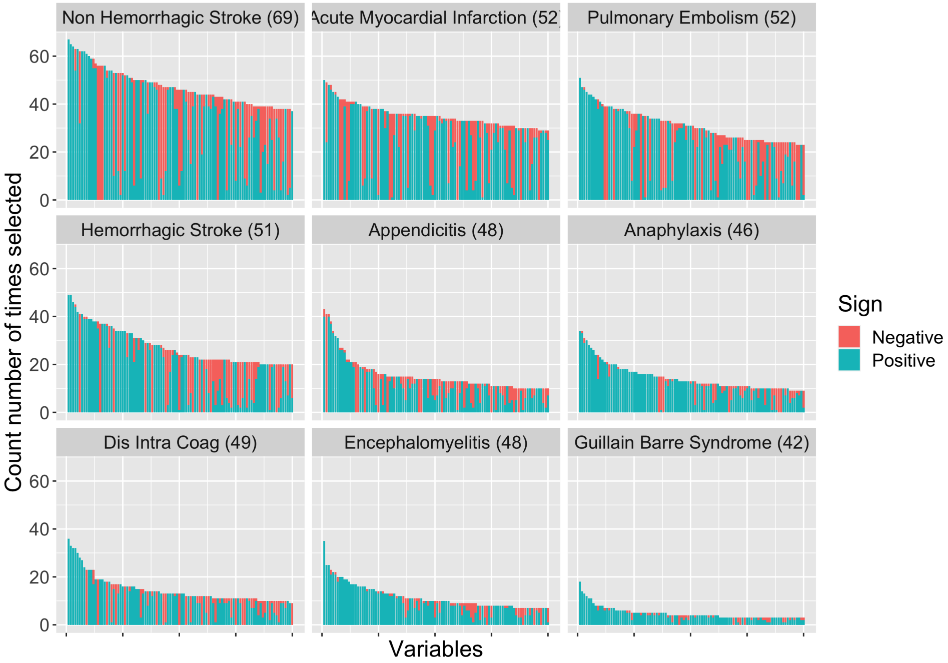

Figure 5: Stacked barplot visualizing the number of times variables are selected and the percentage of times these variables had a positive or negative sign for each prediction task (ordered by a decreasing number of outcomes). The plot is limited to the 100 most frequently selected variables. The number in brackets specifies the number of successfully developed models for each prediction task.

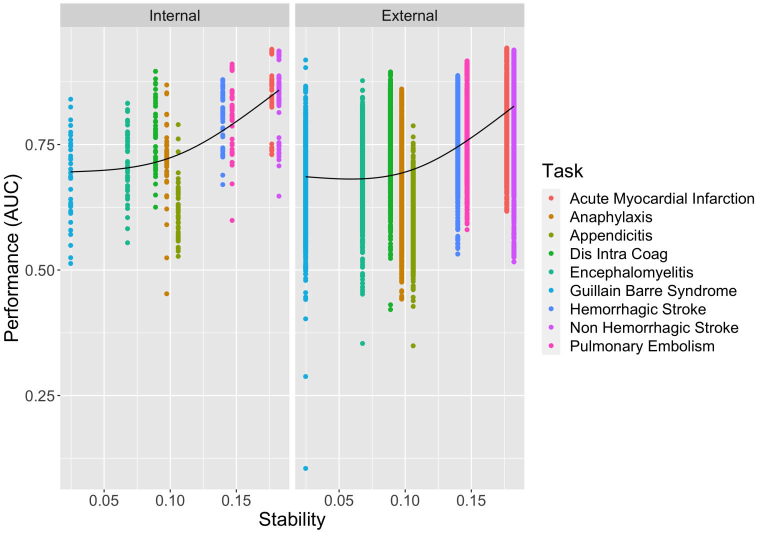

Figure 6: Internal and external predictive performance as measured by the area under the receiver operator curve (AUC) for each prediction task. The black line represents the smoothed conditional means.

5. Discussion

We investigated model stability by training over 450 prediction models. For the nine prediction tasks in this study (corresponding to different COVID-19 vaccine outcomes of interest), that included ∼50 models per task, we found low similarity between models for the same task. This suggests that the developed models are unstable both in terms of the variables included in the model and in the sign of their coefficients. Therefore, the study design choices are likely to impact the results obtained by researchers interpreting LASSO regression to identify 'risk factors'. This is problematic and we recommend either investigating model robustness across settings or using alternative methods for 'risk factor' detection such as univariate analysis.

Model stability is important, but few published prediction models perform sensitivity analyses to investigate the stability of models to the choice of database, target population, and outcome phenotype definition. This study is currently the largest of its kind to investigate model stability.

The results show that model stability, measured in terms of the stability of the selected variables, is poor across the prediction tasks but slightly better for the top (i.e. most important) variables. We found a larger number of outcomes typically leads to larger models. This is in agreement with findings of John et al. (2022), who showed that model size went down when a smaller training sample was used, keeping the outcome proportion constant (i.e. a reduced number of outcomes). Furthermore, we found that a larger number of outcomes is positively correlated with a more stable variable set. For a larger target population, we find the opposite relation (less stability), which might be due to a higher number of available candidate variables. Moreover, we showed differences in the top variables are mostly due to database choice and not due to using different target population and/or outcome phenotype definitions (as long as the number of outcome cases is sufficiently high). This large impact of database choice might be explained by differences in population case-mix or availability of variables between databases, leading to a shift in the top variables.

We showed that direction of effect (i.e. sign) of variables often changes across models; only 27-54% of the variables for different prediction tasks never change sign. This is also true for the top variables (those selected more often) and makes clinical interpretation of the effect of potential 'risk factors' difficult. For LASSO regression (and more generally; variable importance methods) the correlations between the variables are known to impact the coefficients (importance of variables). For example, removing one variable out of a set of correlated variables (that are also highly associated to the outcome) is unlikely to impact a model's performance and will make the removed variable seem less predictive than it is, as one of the other variables is likely to take the place of the removed variable. Unless correlations are accounted for, interpreting a prediction model for 'risk factor' effect is likely to be problematic and we recommend that it should be avoided.

Recently, there is also an increasing interest in explainability; to create insight into how and why models produce predictions (Markus et al., 2021). The results in this paper show that prediction models are unstable even though the discriminative performance is stable. This study highlights it is important to be careful when using LASSO regression to identify 'risk factors' and not to over-interpret the developed models in general. The focus of model evaluation is often internal predictive performance (when the model is used in the development database), with a growing awareness around the importance of external validation (when the model is used in a new clinical setting) (Reps et al., 2020; Yang et al., 2022). However, model stability is often not assessed and we recommend investigating model stability when interpretability of a model is important (e.g. in the case of 'risk factor' detection). Model robustness is also recognized as area for improvement by Goldstein et al. (2017). Alternatively, one can use more traditional methods for 'risk factor' detection such as univariate analysis that do not have the same instability as LASSO regression. However, differences between databases might still persist.

In the future, we hope to extend the experiment to include other algorithms. In particular, it would be interesting to evaluate stability-adjusted versions of LASSO regression (e.g. Meinshausen and Bühlmann (2010)) and other model classes such as tree-based or deep learning methods. Furthermore, it would be interesting to analyze model stability beyond the direction of effect, by comparing the relative importance of selected variables. Finally, investigating to what extent different variables represent the same concept (within or across databases) (Sechidis et al., 2019) and how to best exploit the knowledge in different databases for 'risk factor' detection are promising directions for future research.

Limitations We only investigate LASSO logistic regression and it is unknown whether the results will generalize to other algorithms. Furthermore, this is a case study investigating only a limited number of target population and outcome phenotype definitions and results may depend on the selected prediction tasks. For example, model stability results might be inaccurate if very poor phenotype definitions are used. However, evaluating phenotype definitions for observational data is difficult in practice as there is typically no ground truth (e.g., in claims data chart review is not possible). Results may further depend on the complexity of the included prediction tasks. Finally, we did not investigate within-sample stability (same database, same definitions), as we were not interested to isolate the effect of study design choices, but rather to investigate the overall impact of different choices on the stability of models.

Acknowledgments

The authors are grateful for the valuable comments of Dr. Patrick B. Ryan that helped shape and improve this manuscript.

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.

References

Evangelia Christodoulou, Jie Ma, Gary S Collins, Ewout W Steyerberg, Jan Y Verbakel, and Ben Van Calster. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, 110:12–22, 2019. doi: 10.1016/j.jclinepi.2019.02.004.

Chad A Davis, Fabian Gerick, Volker Hintermair, Caroline C Friedel, Katrin Fundel, Robert Küffner, and Ralf Zimmer. Reliable gene signatures for microarray classification: assessment of stability and performance. *Bioinformatics*, 22(19):2356–2363, 2006.

Kevin Dunne, Padraig Cunningham, and Francisco Azuaje. Solutions to instability problems with sequential wrapper-based approaches to feature selection. *Journal of Machine Learning Research*, 1:22, 2002.

Ángel García de la Garza, Carlos Blanco, Mark Olfson, and Melanie M Wall. Identification of Suicide Attempt Risk Factors in a National US Survey Using Machine Learning. *JAMA Psychiatry*, 78(4):398–406, 04 2021. ISSN 2168-622X. doi: 10.1001/jamapsychiatry.2020.4165.

Wilson Wen Bin Goh and Limsoon Wong. Evaluating feature-selection stability in next-generation proteomics. *Journal of Bioinformatics and Computational Biology*, 14(05): 1650029, 2016.

Benjamin A Goldstein, Ann Marie Navar, Michael J Pencina, and John Ioannidis. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 24 (1):198–208, 2017.

Shivapratap Gopakumar. *Machine learning in healthcare: an investigation into model stability*. PhD thesis, Deakin University, 2017.

Jill Hardin and Jenna M Reps. Evaluating the impact of covariate lookback times on performance of patient-level prediction models. *BMC medical research methodology*, 21 (1):1–9, 2021.

Luis H. John, Jan A. Kors, Jenna M. Reps, Patrick B. Ryan, and Peter R. Rijnbeek. Logistic regression models for patient-level prediction based on massive observational data: Do we need all data? *International Journal of Medical Informatics*, 163:104762, 2022. ISSN 1386-5056. doi: 10.1016/j.ijmedinf.2022.104762.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms. In *Fifth IEEE International Conference on Data Mining (ICDM'05)*, pages 218–225. IEEE, 2005. doi: 10.1109/ICDM.2005.135.

Sara Khalid, Cynthia Yang, Clair Blacketer, 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*, 211:106394, 2021. doi: 10.1016/j.cmpb.2021.106394.

Pavel Křížek, Josef Kittler, and Václav Hlaváč. Improving stability of feature selection methods. In *International Conference on Computer Analysis of Images and Patterns*, pages 929–936. Springer, 2007.

Chenlei Leng, Yi Lin, and Grace Wahba. A note on the lasso and related procedures in model selection. *Statistica Sinica*, pages 1273–1284, 2006.

Aniek F Markus, Jan A Kors, and Peter R Rijnbeek. The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics*, 113: 103655, 2021.

Archana J McEligot, Valerie Poynor, Rishabh Sharma, and Anand Panangadan. Logistic lasso regression for dietary intakes and breast cancer. *Nutrients*, 12(9):2652, 2020.

Nicolai Meinshausen and Peter Bühlmann. Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso. Seminar für Statistik, Eidgenössische Technische Hochschule (ETH), Zürich, 2004.

Nicolai Meinshausen and Peter Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4):417–473, 2010.

Robert J Mentz, L Kristin Newby, Ben Neely, et al. Assessment of administrative data to identify acute myocardial infarction in electronic health records. *Journal of the American College of Cardiology*, 67(20):2441–2442, 2016. ISSN 0735-1097. doi: 10.1016/j.jacc.2016.03.511.

Sarah Nogueira, Konstantinos Sechidis, and Gavin Brown. On the stability of feature selection algorithms. *Journal of Machine Learning Research*, 18(174):1–54, 2018. doi: 10.5555/3122009.3242031.

Simon Nusinovici, Liang Zhang, Xiaoran Chai, Lei Zhou, Yih Chung Tham, Caroline Vasseneix, Shivani Majithia, Charumathi Sabanayagam, Tien Yin Wong, and Ching Yu Cheng. Machine learning to determine relative contribution of modifiable and non-modifiable risk factors of major eye diseases. *British Journal of Ophthalmology*, 106 (2):267–274, 2022.

J Marc Overhage, Patrick B Ryan, Christian G Reich, Abraham G Hartzema, and Paul E Stang. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60, 2012.

Arjun Parthipan, Imon Banerjee, Keith Humphreys, Steven M Asch, Catherine Curtin, Ian Carroll, and Tina Hernandez-Boussard. Predicting inadequate postoperative pain management in depressed patients: a machine learning approach. *PLoS One*, 14(2): e0210575, 2019.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3:96–146, 2009.

Bhargava K Reddy, Dursun Delen, and Rupesh K Agrawal. Predicting and explaining inflammation in crohn's disease patients using predictive analytics methods and electronic medical record data. *Health Informatics Journal*, 25(4):1201–1218, 2019.

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*, 25(8):969–975, 2018. doi: 10.1093/jamia/ocy032.

Jenna M Reps, Ross D Williams, Seng Chan You, Thomas Falconer, Evan Minty, Alison Callahan, Patrick B Ryan, Rae Woong Park, Hong-Seok Lim, and Peter Rijnbeek. 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 Medical Research Methodology*, 20(1):1–10, 2020.

C Mary Schooling and Heidi E Jones. Clarifying questions about "risk factors": predictors versus explanation. *Emerging Themes in Epidemiology*, 15(1):1–6, 2018.

Konstantinos Sechidis, Konstantinos Papangelou, Sarah Nogueira, James Weatherall, and Gavin Brown. On the stability of feature selection in the presence of feature correlations. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 327–342. Springer, 2019.

L Shi, LH Reid, WD Jones, and others. The microarray quality control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151, 2006.

Marc A Suchard, Shawn E Simpson, Ivan Zorych, Patrick Ryan, and David Madigan. Massive parallelization of serial inference algorithms for complex generalized linear models. *ACM Transactions on Modeling and Computer Simulation*, 23:10, 2013. doi: 10.1145/2414416.2414791.

U.S. Food and Drug Administration. FDA Updates of COVID-19 Vaccine Safety Activities. 2021. URL https://www.fda.gov/media/150051/download. Accessed on 2022-04-01.

Sijian Wang, Bin Nan, Saharon Rosset, and Ji Zhu. Random lasso. *The Annals of Applied Statistics*, 5(1):468, 2011.

Cynthia Yang, Jan A Kors, Solomon Ioannou, Luis H John, Aniek F Markus, Alexandros Rekkas, Maria A J de Ridder, Tom M Seinen, Ross D Williams, and Peter R Rijnbeek. Trends in the conduct and reporting of clinical prediction model development and validation: a systematic review. *Journal of the American Medical Informatics Association*, 01 2022. doi: 10.1093/jamia/ocac002. ocac002.

Sheng Zhang, Sisi Huang, Jiao Liu, Xuan Dong, Mei Meng, Limin Chen, Zhenliang Wen, Lidi Zhang, Yizhu Chen, Hangxiang Du, Yongan Liu, Tao Wang, and Dechang Chen. Identification and validation of prognostic factors in patients with covid-19: A retrospective study based on artificial intelligence algorithms. *Journal of Intensive Medicine*, 1(2): 103–109, 2021. ISSN 2667-100X. doi: 10.1016/j.jointm.2021.04.001.

Appendix A. Database descriptions

- IBM MarketScan Commercial Claims and Encounters (CCAE) database represents data from individuals enrolled in United States employer-sponsored insurance health plans. The data includes adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy) as well as enrollment data from large employers and health plans who provide private healthcare coverage to employees, their spouses, and dependents. Additionally, it captures laboratory tests for a subset of the covered lives. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.

- Optum® De-identified Electronic Health Record Dataset (Optum EHR) is a multidimensional database containing information on outpatient visits, diagnostic procedures, medications, laboratory results, hospitalizations, clinical notes and patient outcomes primarily from IDNs. The EHR database includes representation of 80M patients with at least 7M patients from each US Census region. Furthermore, the database contains a provider network of 140,000+ providers at more than 700 hospitals and 7,000 clinics. This database does not have eligibility controls but researchers can track patient activity. More than 45% of patients have activity well over 3 years and more than 30% of patients have activity spanning over 5 years.

- Optum® De-Identified Clinformatics® Data Mart (Optum DoD) is an adjudicated US administrative health claims database for members of private health insurance, who are fully insured in commercial plans or in administrative services only (ASOs), Legacy Medicare Choice Lives (prior to January 2006), and Medicare Advantage (Medicare Advantage Prescription Drug coverage starting January 2006). The population is primarily representative of commercial claims patients (0-65 years old) with some Medicare (65+ years old) however ages are capped at 90 years. It includes data captured from administrative claims processed from inpatient and outpatient medical services and prescriptions as dispensed, as well as results for outpatient lab tests processed by large national lab vendors who participate in data exchange with Optum. This dataset also provides date of death (month and year only) for members with both medical and pharmacy coverage from the Social Security Death Master File (however after 2011 reporting frequency changed due to changes in reporting requirements) and location information for patients is at the US state level.

- IBM MarketScan Multi-State Medicaid (MDCD) database 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 and Medicare eligibility. Members maintain their same identifier even if they leave the system for a brief period however the dataset lacks lab data.

- IQVIA Disease Analyzer Germany (IQVIA Germany) database consists of data collected from physician practices and medical centers for all ages. Mostly primary care physician data however some data from specialty practices (where practices are electronically connected to each other) and some lab data is included. Key attributes include demographics, prescriptions as prescribed at brand level, diagnosis, lab measurements, actions (e.g. referrals, sick notes).

- Japan Medical Data Center (JMDC) database is a payer based database that has collected claims, ledger of the insured people and health checkup results from more than 250 payers. It covers workers and their dependents aged under 74. It is longitudinal and the largest one as commercially available database in Japan with more than 13 million enrollments. All medical history of the insured people are available and patient reported outcome research can be done through payers on-demand basis. Those aged 66 or older are less representative as compared with whole population in the nation. When estimated among the people who are younger than 66 years old, the proportion of children younger than 18 years old in JMDC is approximately the same as the proportion in the whole nation. Claims data are derived from monthly claims issued by clinics, hospitals and community pharmacies. The number of claims issued and added to JMDC database is about 6,000,000 per month. The size of JMDC population is about 6% of the whole nation.

- IBM MarketScan Medicare Supplemental (MDCR) database 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. These data include adjudicated health insurance claims (e.g. inpatient, outpatient, and outpatient pharmacy). Additionally, it captures laboratory tests for a subset of the covered lives.

Appendix B. Phenotype definitions

Table 2: List of phenotype definitions used for each outcome. Abbreviations: 'Broad' = broad definition including a wide range of diagnosis codes, 'Narrow' = narrow definition including a smaller set of diagnosis codes selected by the U.S. Food and Drug Administration, 'IP' = restricted to inpatient visits, 'ED' = restricted to emergency department visits.

Outcome	**Phenotype**	**URL**
Acute myocardial infarction	Broad	https://atlas.ohdsi.org/#/cohortdefinition/383
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/340
	Narrow + IP	https://atlas.ohdsi.org/#/cohortdefinition/388
Anaphylaxis	Broad	https://atlas.ohdsi.org/#/cohortdefinition/349
	Broad + IP/ED	https://atlas.ohdsi.org/#/cohortdefinition/407
	Narrow	https://atlas.ohdsi.org/#/cohortdefinition/389
Appendicitis	Broad	https://atlas.ohdsi.org/#/cohortdefinition/386
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/344
	Narrow	https://atlas.ohdsi.org/#/cohortdefinition/390
Disseminated intravascular coagulation	Broad	https://atlas.ohdsi.org/#/cohortdefinition/385
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/336
	Narrow	https://atlas.ohdsi.org/#/cohortdefinition/392
Encephalomyelitis	Broad	https://atlas.ohdsi.org/#/cohortdefinition/382
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/346
	Narrow + IP	https://atlas.ohdsi.org/#/cohortdefinition/393
Guillain-Barré Syndrome	Broad	https://atlas.ohdsi.org/#/cohortdefinition/380
	Broad + IP/ER	https://atlas.ohdsi.org/#/cohortdefinition/343
	Broad + IP/ER (primary condition)	https://atlas.ohdsi.org/#/cohortdefinition/348
Hemorrhagic stroke	Broad	https://atlas.ohdsi.org/#/cohortdefinition/387
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/341
	Narrow + IP	https://atlas.ohdsi.org/#/cohortdefinition/405
Non-hemorrhagic stroke	Narrow	https://atlas.ohdsi.org/#/cohortdefinition/408
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/406
	Narrow + IP	https://atlas.ohdsi.org/#/cohortdefinition/397
	Broad	https://atlas.ohdsi.org/#/cohortdefinition/384
Pulmonary embolism	Broad	https://atlas.ohdsi.org/#/cohortdefinition/411
	Broad + IP	https://atlas.ohdsi.org/#/cohortdefinition/404
	Narrow	https://atlas.ohdsi.org/#/cohortdefinition/400

Appendix C. Additional results

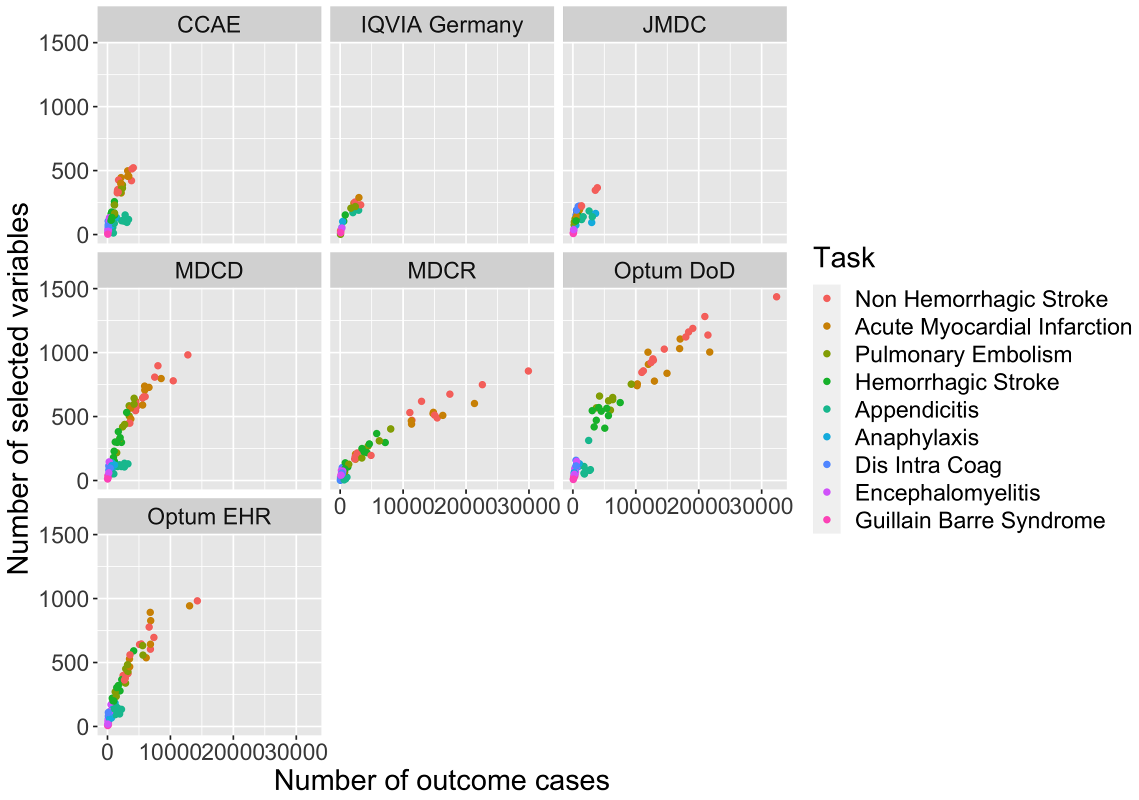

Figure 7: Scatterplot of the number of outcomes versus the number of selected variables per database. Each point represents a developed model, colors indicate the corresponding prediction task.

Table 3: List of top 10 most often selected variables across prediction tasks (variable from condition domain* or drug domain †, positively (+) or negatively (-) contributing to outcome risk in majority of models).

Task	**Variables**
Acute myocardial infarction	Heart disease*(+), Age 65-69(+/-), Vascular disorder*(+), Male(+), Myocardial disease*(+), Platelet aggregation inhibitors excl. heparin†(+), Ischemic heart disease*(+), Age 10-14(-), Age 15-19(-), Hyptertensive disorder*(+)

Anaphylaxis	Asthma*(+), Adrenergics, inhalents†(+), Traumatic injury*(+), Cardiac therapy†(+), Age 15-19(+), Blood and blood forming organs†(+), Edema*(+), Allergic disposition*(+), Male(+), Epinephrine†(+)
Appendicitis	Male(+), Age 0-4(-), Abdominal pain*(+), Age 10-14(+), Age 15-19(+), Pain of truncal structure*(+), Age 20-24(+), Age 25-29(+), Pain finding at anatomical site*(+), Inflammatory disorder of digestive tract*(+)
Disseminated intravascular coagulation	Heart disease*(+), Abnormal blood cell count*(+), Vascular disorder*(+), Blood and blood forming organs†(+), Measurement finding outside reference range*(+), Antineoplastic and immunomodulating agents†(+), Antithrombotic agents†(+), Complication of procedure*(+), Hyperlipidemia*(-), Kidney disease*(+)
Encephalomyelitis	Antiepileptics†(+), Measurement finding outside reference range*(+), Male(+), Fatigue(+), Vascular disorder*(+), Inflammation of specific body systems*(+), Muscle weakness*(+), Nervous system†(+), Abnormal blood cell count*(+), Neuropathy*(+)
Guillain-Barré syndrome	Neuropathy*(+), Peripheral nerve disease*(+), Essential hypertension*(+), Fatigue*(+), Vascular disorder*(+), Male(+), General problem and/or complaint*(+), Arthropathy*(+), Measurement finding above reference range*(+), Polyneuropathy*(+)
Hemorrhagic stroke	Bleeding*(+), Lesion of brain*(+), Vascular disorder*(+), Male(+), Seizure*(+), Age 10-14(-), Antiepileptics†(+), Antithrombotic agents†(+), Blood and blood forming organs†(+), Drug dependence*(+)
Non-hemorrhagic stroke	Cerebrovascular disease*(+), Hyptertensive disorder*(+), Heart disease*(+), Cerbral infarction*(+), Lesion of brain*(+), Age 65-69(+/-), Headache*(+), Antithrombotic agents†(+), Blood and blood forming organs†(+), Vascular disorder*(+)

Pulmonary embolism	Vascular disorder*(+), Antithrombotic agents†(+), Blood and blood forming organs†(+), Embolism*(+), Obesity*(+), Secondary malignant neoplastic disease*(+), Soft tissue lesion*(+), Primary malignant neoplasm*(+), Male(+), Heart disease*(+)

Table 4: The percentage of variables with the same sign (relative to all variables that occur in at least three models) for each prediction task.

Task	%
Acute myocardial infarction	28
Anaphylaxis	45
Appendicitis	27
Disseminated intravascular coagulation	54
Encephalomyelitis	40
Guillain-Barré syndrome	44
Hemorrhagic stroke	36
Non-hemorrhagic stroke	27
Pulmonary embolism	41