Supervisor: Martin Gerdin Wärnberg

**A Segmented Approach to Predicting Prediction Model Performance After Transfer Using Unlabeled Data**

Seminar version

**Author: Arman Norouzi**

Department of Global Public Health  
Arman Norouzi  
Study Program in Medicine KI  
Degree project 30 credits  
Fall 2020



****

**Ett segmenterat tillvägagångssätt med omärkt data för att förutsäga prestanda hos en prediktionsmodell efter överföring till ny kontext**

*Bakgrund:* De flesta prediktionsmodeller som implementerats inom hälso- och sjukvården presterar sämre när de överförts till andra länder eller kontexter som skiljer sig från den kontext där de utvecklats. För närvarande finns det ingen metod som kan förutsäga prestandan av prediktionsmodeller efter att de överförts genom att använda omärkt data från överförningslandet, varför vi har utvecklat och testat en sådan metod. *Syfte*: Att utveckla och testa en metod för att förutsäga prestanda av en prediktionsmodell efter överföring till ett annat land med omärkt data. *Material och Metoder:* Vi använde oss av publikt tillgänglig data med deltagare från tre länder för att simulera data. Denna data användes för att utveckla prediktionsmodeller i varje land som sedan användes för att bedöma noggranheten hos modellerna inom det land som de utveckalts i, i det land som de överförts till samt i ett segmenterat urval från det land där de utvecklats i. Dessa noggrannheter jämfördes och hela processen upprepades 1000 gånger för att beräkna konfidensintervall. *Resultat*: Vi fann i två av våra sex överförningskombinationer att vår metods förutsagda noggrannhet var signifikant bättre på att förutsäga noggrannheten efter att prediktionsmodellen överförts jämfört med noggrannheten i utvecklingslandet. I de återstående fyra överförningskombinationerna fann vi resultat som var exakt tvärtom. *Slutsats*: Våra resultat stödjer inte användningen av vår metod för att förutsäga noggrannheten av en prediktionsmodell efter att den överförts till ett annat land. Ytterligare studier krävs dock på grund av de begränsningar som fanns i vår studie.

**A segmented approach to predicting prediction model performance after transfer using unlabeled data**

*Introduction:* Most prediction models implemented in health care perform worse when they are transferred to another country that differs from the country in which the prediction model was developed. At present, there are no methods that can predict the performance of a prediction model after transfer using only unlabeled data from the transfer country. Therefore, we have developed and tested such a method. *Aims:* To develop and test a method that predicts prediction model performance after transfer using unlabeled data. *Material and Methods:* We used a public dataset with participant data from three countries to simulate new samples in each country. These samples were used to develop prediction models in each country in order to assess the accuracy of those models within the country that they were developed in, in the country that they were transferred to and in a segment of the sample within the country that the model was developed in. These accuracies were compared and the whole process was repeated 1000 times to develop confidence intervals. *Results:* We found that in two of our six transfer combinations that our method’s predicted accuracy was significantly better than the accuracy in the development country at predicting the accuracy after transfer. In the remaining four transfer combinations, we found results that were the exact opposite. *Conclusions:* Our results do not support the use of our method to predict performance after transfer. However, further studies are required due to limitations of our study.

*Keywords:* Prediction models; Diagnostic models; Prognostic models; Logistic regression; External validation.

# Introduction

## Prediction models

In medicine, health care professionals are confronted with a wide range of information that needs to be processed in order to make informed clinical decisions. To help health care professionals make such decisions, prediction models (also referred to as prediction scores or prediction rules) have been implemented in health care (1, 2). These prediction models can be defined as statistical algorithms that predict the risk of a specific outcome occurring in an individual based on predictors (2, 3). They are capable of such tasks due to being trained to identify patterns in predictor data that have been labeled with an outcome of interest. These patterns can then be used to predict outcome based on new unlabeled predictor data (4).

### Diagnostic models and prognostic models

The risk that is predicted by the prediction models is generally based on multiple predictors and the outcome could either be a disease (diagnostic model) or an event that will occur in the future (prognostic model) (3, 5). In a diagnostic model, the predicted risk can be used to reassure the patient that their symptoms are not caused by a serious disease, refer the patient to further testing or to initiate treatment (3). An example of a diagnostic model is the Ottawa Ankle Rules. This prediction model helps predict the risk of a fracture in patients with acute ankle injuries. To predict this risk, the model uses predictor data such as bone tenderness at different locations and the inability to bear weight on the injured foot immediately after injury and in the emergency department (ED). Based on the predicted risk, the healthcare professionals can decide whether the patient needs x-ray imaging (6).

In a prognostic model, the predicted risk can be used to choose between therapeutic options, plan lifestyle changes and to risk-stratify patients in therapeutic clinical trials (2, 7–9). An example of a prognostic model is the CHA2DS2-VASc score. This prediction model helps health care professionals by predicting the annual risk of developing an ischemic stroke in patients with atrial fibrillation. To predict this risk, the prediction model uses predictor data such as history for congestive heart failure, hypertension, age >74, diabetes, stroke/transient ischemic attack/thromboembolism, vascular disease, age 65-74 and female sex (10). Based on the predicted risk, health care professionals can decide whether a patient needs anticoagulation treatment (11).

## Prediction model studies

There are many uses for prediction models within the fields of medicine, where the Ottawa Ankle Rules and the CHA2DS-VASc score are just two examples that have been implemented in clinical practice. To develop and implement such useful models within health care, several steps are needed to be carried out. These steps include model development studies, model validation studies and model impact studies (12, 13).

### Prediction model development study

In the first step consisting of the model development study, the aim is to develop a prediction model (12). The prediction model is developed by applying a development sample to a statistical algorithm. There are many algorithms to choose from, but usually when the development sample is small, a simpler algorithm is utilized such as logistic regression (4). The development sample consists of predictor data labeled with relevant outcome, which is used to train the algorithm in finding patterns between the predictors and the outcomes (4, 12). When the prediction model has been developed, it usually tends to be optimistic in its performance within the development sample (14). It is therefore important to quantify such optimism through internal validation techniques (7). The quantified optimism can thereafter be adjusted for by applying shrinkage or penalization to the prediction model (15).

### Prediction model validation study

In the second step consisting of the model validation study, the aim is to assess the performance of the prediction model within a validation sample. The validation sample consists of new individuals with outcome labeled predictor data that differ in various ways from the individuals in the development sample. These individuals may differ in the time in which their data were collected (temporal validation) or which country or hospital their data were collected (geographical validation). Geographical validation assesses the transportability of the prediction model. Such external validation is important as most internally validated prediction models perform worse when applied to new individuals that differ from those used to train the prediction model (13).

### Prediction model impact study

In the third and final step consisting of the model impact study, the aim is to assess the prediction models impact, ideally in a randomized trial. The impact of the model is assessed in variables such as decision-making changes in health care professionals, patient health outcomes or cost-effectiveness of care. These impact studies are carried out to prove that the prediction model is of value in clinical practice (13).

## Problem that can occur during prediction model studies

Carrying out these prediction model studies can be complex as problems may arise during the time in which they are carried out. One such problem may occur during the model validation studies. In these studies, in order to obtain the performance of the prediction model within the validation sample, both predictor and outcome data is required from the validation sample (13). This data is not always available retrospectively and can present a problem that is both time inefficient and expensive, if the data is difficult to access when collecting it prospectively.

This would for example be a problem if data for the Framingham Risk Score were to be collected prospectively to perform a model validation study. The predictor data in this prediction model are cheap blood samples and simple demographics while the outcome data is cardiovascular disease within 10 years (16). The predictor data for this prediction model may be easily accessible but the outcome data is only accessible after 10 years of follow up.

It would therefore be desirable to have a method that can predict the performance of a prediction model after transfer by using unlabeled predictor data from the validation sample. Such a method could in theory, simplify the process of implementing prediction models in clinical practice and therefore indirectly improve decision-making changes in health care professionals, patient health outcomes or cost-effectiveness of care. At present, no such studies nor methods exists which presents a substantial knowledge gap. Therefore, the aim of this study was to develop and test a new method that predicts prediction model performance after transfer using unlabeled data.

### Hypothesis

Our hypothesis was that our method’s predicted accuracy would be as good or better at predicting the accuracy of the prediction model after transfer when compared with the accuracy of the prediction model within the country it was developed in. This hypothesis was because the method’s predicted accuracy would be derived from a segment of the development sample. This segment should theoretically have more similar distribution of model predictors as in the validation sample, hence the hypothesis.

# Aim

The aim of this study was to develop and test a new method for predicting prediction model performance after transfer using unlabeled data.

# Methods and Materials

## Study design

The study design was a simulation study. To perform the analysis, a dataset from a multinational observational study that has been made freely reusable by Eckert A et al in the Dryad Digital Repository was used (17, 18). This dataset was chosen due to consisting of participant data from three different countries with available patient parameters that can be associated to a patient outcome.

## Participants

The participants enrolled in the dataset were all patients seeking ED care between March 2013 and October 2014 within three tertiary care centers in the USA (Clearwater Hospital), France (Hôspital de la Salpêtrière) and Switzerland (Kontonsspital Aaura). The data that was registered for each participant included the hospital and the country in which the patient sought ED care, vital signs, laboratory assessments, age, discharge location, length of stay, intensive care unit (ICU) admission and death within 30 days. The inclusion criteria to be enrolled in the dataset was that an initial blood sample was taken. The exclusion criteria were pediatric or surgical patient (17).

## Variables

### Model predictors

The model predictors that were used from the dataset in order to develop prediction models and simulate new model predictors in the statistical analysis, were respiratory rate (per min), peripheral oxygen saturation (%), systolic blood pressure (mm Hg), heart rate (bpm), temperature (°C) and age. Although there is no consensus on which method is best for selecting model predictors, it has been recommended that all available model predictors should be included to reduce overfitting and selection bias (12). Due to simplicity, we chose to include all model predictors that were continuous and that had no missing data. How these model predictors were measured was not mentioned in the study that publicized them.

### Model outcomes

The model outcomes that were used from the dataset in order to develop prediction models and simulate new model outcomes in the statistical analysis included ICU admission and the country from which the patient sought ED care. The decision to admit the patients to the ICU was left to the treating physician. ICU admission was chosen as the outcome for the prediction model due being more frequent than death within 30 days.

### Sample size

The final sample size used in this study was 1303 participants which included all the participants from the dataset.

### Missing data

Because of the dataset already being filtered to mostly containing no missing data, a complete case analysis was carried out.

## Statistical analysis

### Dataset

The dataset previously mentioned in the study design was divided based on the country from which the participants sought ED care (USA sample, France sample and Switzerland sample).

### Sequence of analysis

Analysis in this study was performed in the programming language R (19). The decision threshold for all the prediction models were set to 0.5. The sequence of analysis performed were sample simulation, sample assignment, prediction model development, development sample accuracy, validation sample accuracy, propensity model development, predicted validation sample accuracy and approach comparison.

### Sample simulation

To increase the number of participants, 10000 new participants were simulated for each of the divided samples. The process of simulation included a model predictor simulation and a model outcome simulation. The model predictors that were used to simulate new model predictors were respiratory rate, peripheral oxygen saturation, systolic blood pressure, heart rate, temperature and age from the divided samples. The model outcome that was used to simulate new model outcomes were ICU admission from the divided samples.

To perform the model predictor simulation for one of the divided samples, the mvrnorm function implemented in the MASS package was used (20). The function used the mean and the covariance of the model predictors in the divided sample to simulate new model predictors. To perform the model outcome simulation for the newly simulated model predictors, the glm function implemented in R was used to develop a logistic regression model. This model was trained with the model predictors and model outcomes from the divided sample. The model was then used to predict outcomes in the newly simulated model predictors. These predictions were set as the outcomes for the newly simulated model predictors. The simulated model predictors with its model outcomes constituted a simulated sample from one country. The model predictor simulation and model outcome simulation process were repeated until each divided sample had a simulated sample.

### Sample assignment

To simulate the transfer of a prediction model from one country to another, one of the simulated samples was denoted as the development sample while one of the two remaining simulated samples was denoted as the validation sample. The development sample represented data from the country in which the prediction model was created, while the validation sample represented data from the country in which the prediction model was transferred to.

### Prediction model development

In the prediction model development step, a prediction model was developed by training a logistic regression model with the development sample. To reduce the error difference between the development sample accuracy and the validation sample accuracy, we used cross-validation to choose the model within one standard error from the best model. This was done using the caret package implemented in R (21). The cross-validation technique was based on five folds. The model predictors that were used to train the model were respiratory rate, peripheral oxygen saturation, systolic blood pressure, heart rate, temperature, and age. The model outcome that was used to train the model included ICU admission.

### Development sample accuracy

To assess the accuracy of the prediction model within the country that it was developed in, the model developed in the prediction model development step was used to predict outcome within the development sample. The predictions where then compared with the true outcomes in the development sample in order to acquire the development sample accuracy.

### Validation sample accuracy

To assess the accuracy of the prediction model within the country to which it was transferred to, the model developed in the prediction model development step was used to predict outcome within the validation sample. The predictions where then compared with the true outcomes in the validation sample in order to acquire the validation sample accuracy.

### Propensity model development

In the propensity model development step, the data from the development sample and the validation sample were pooled into one sample. This aggregated sample was used to develop a propensity model, also a prediction model, by training a logistic regression model with the aggregated sample. The propensity model was then used to predict the origin of the samples in the aggregated sample. Observations from the development sample that were misclassified as validation observations, were used to identify a segment of the development sample with observations similar to the validation sample. The model predictors that were used to train the propensity model were respiratory rate, peripheral oxygen saturation, systolic blood pressure, heart rate, temperature, and age. The model outcome that was used to train the propensity model was the country in which the participant sought ED care.

### Predicted validation sample accuracy

To assess our method’s predicted accuracy of the prediction model within the country it was transferred to, the model developed in the prediction model development step was used to predict outcome within the segmented sample created in the propensity model development step. The predictions where then compared with the true outcomes in the segmented sample in order acquire the predicted validation sample accuracy.

### Approach comparison

To assess the error in the “naive approach”, the difference between the development sample accuracy and the validation sample accuracy was calculated. To assess the error in the “segmented approach”, the difference between our method’s predicted validation sample accuracy and the validation sample accuracy was calculated. To assess which approach performed best, the difference between the absolute value of the naive approach and the absolute value of the segmented approach was calculated.

### Sequence repetition

To obtain 95% confidence intervals (CI) around the accuracies and the differences, the sequence of analysis was repeated 1000 times. These repetitions were performed for each available combination in the sample assignment step.

## Ethical considerations

### Principle of autonomy

The dataset that was used in this study has been made freely reusable in Dryad Digital Repository (18). Therefore, the principle of autonomy is upheld due to there not being any requirement for informed consent.

### Principle of beneficence

This study attempted to act in the best interest of future analytical research and patients, by developing and testing a new method for predicting prediction model performance after transfer using unlabeled data. Such a method could in theory simplify the process of implementing prediction models in clinical practice and therefore indirectly improve decision-making changes in health care professionals, patient health outcomes or cost-effectiveness of care.

### Principle of nonmaleficence

The method developed in this study will be made without the intention of harm, intentionally or unintentionally. To nullify the risk of patient identification leakage, we used a dataset that has already been depersonalized and made freely reusable. By taking these actions we determined that the risk to the population is minimal.

### Principle of justice

Due to this study being analytical, the principle of justice does not prevail. However, the data in the study was treated equally.

### Ethical permit

No ethical permit was required because this study used a public database.

# Results

## Original Sample description

All 1303 participants (USA 940, France 355, Switzerland 8) in the original dataset were used to simulated new participants. 1132 out of these participants were not admitted to the ICU and their mean age was 64.73 years with 50.5 % being males. Baseline characteristics of the original dataset stratified by country are shown in table 1.

**Table 1: Baseline characteristics of the original dataset stratified by country.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Overall | France | Switzerland | USA |
| n | 1303 | 355 | 8 | 940 |
| Respiratory rate (per min) (mean (SD)) | 19.92 (5.24) | 22.30 (5.95) | 31.25 (7.92) | 18.93 (4.47) |
| Gender = Male (%) | 658 (50.5) | 182 (51.3) | 3 (37.5) | 473 (50.3) |
| Peripheral oxygen saturation (%) (mean (SD)) | 96.42 (3.87) | 96.48 (3.65) | 90.62 (10.56) | 96.45 (3.82) |
| Systolic blood pressure (mm Hg) (mean (SD)) | 141.26 (30.16) | 134.71 (23.82) | 147.25 (21.49) | 143.68 (31.96) |
| Pulse (bpm) (mean (SD)) | 85.70 (21.27) | 87.95 (20.90) | 88.25 (25.19) | 84.83 (21.34) |
| Temperature (°C) (mean (SD)) | 36.58 (0.70) | 36.75 (0.68) | 37.49 (1.37) | 36.51 (0.68) |
| Age (mean (SD)) | 64.73 (18.01) | 58.40 (19.29) | 66.00 (9.83) | 67.11 (16.97) |
| ICU admission = No admission (%) | 1132 (86.9) | 318 (89.6) | 7 (87.5) | 807 (85.9) |

## Simulation sample description

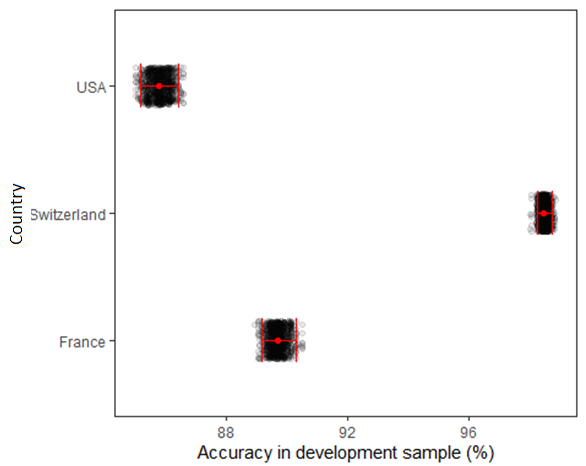
With the 1303 participants in the original dataset, 10000000 new participants were simulated for each country. The mean age of the participants in the simulated sample was 63.84 years. Baseline characteristics of the simulated samples stratified by country are shown in table 2.

**Table 2: Baseline characteristics of the simulated samples stratified by country.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Overall | France | Switzerland | USA |
| n | 30000000 | 10000000 | 10000000 | 10000000 |
| Age (mean (SD)) | 63.84 (16.35) | 58.40 (19.29) | 66.00 (9.83) | 67.11 (16.97) |
| Respiratory rate (per min) (mean (SD)) | 24.16 (8.15) | 22.30 (5.95) | 31.25 (7.92) | 18.93 (4.47) |
| Peripheral oxygen saturation (%) (mean (SD)) | 94.52 (7.35) | 96.48 (3.65) | 90.62 (10.55) | 96.45 (3.82) |
| Systolic blood pressure (mm Hg) (mean (SD)) | 141.88 (26.67) | 134.71 (23.82) | 147.25 (21.48) | 143.68 (31.96) |
| Pulse (bpm) (mean (SD)) | 87.01 (22.61) | 87.95 (20.90) | 88.25 (25.19) | 84.83 (21.34) |
| Temperature (°C) (mean (SD)) | 36.92 (1.05) | 36.75 (0.68) | 37.49 (1.37) | 36.51 (0.68) |
| ICU admission = No admission (%) | 26045048 (86.8) | 8968011 (89.7) | 8499784 (85.0) | 8577253 (85.8) |

## Development sample accuracies

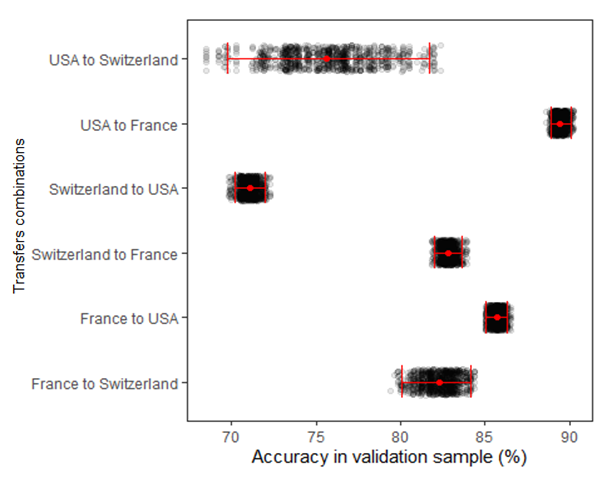
We found the mean development sample accuracy to be 85.78 % (95% CI 85.17 to 86.39) within USA, 98.48 % (95% CI 98.26 to 98.76) within Switzerland and 89.68 % (95% CI 89.16 to 90.27) within France. All development sample accuracies with mean accuracies and CIs are shown in figure 1 stratified by country.



**Figure 1: Development sample accuracies in each country.** Each black dot represents the development sample accuracy for one simulated sample with the red dot representing the mean accuracy. The bars indicate the 95% CI around the development sample accuracies.

## Validation sample accuracies

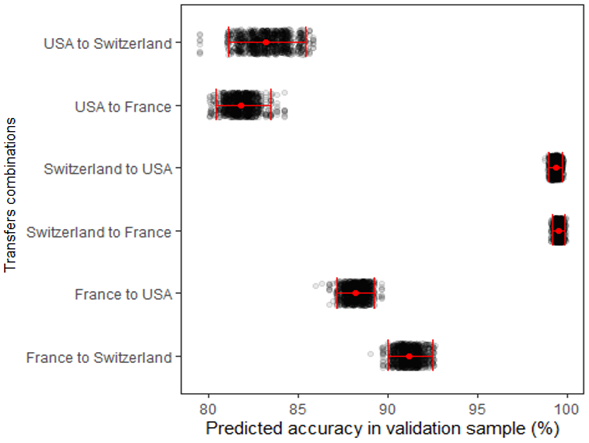
We found the mean validation sample accuracy to be 75.59 % (95% CI 69.72 to 81.71) within the USA to Switzerland transfer, 89.46 % (95% CI 88.91 to 90.08) within the USA to France transfer, 71.05 % (95% CI 70.19 to 71.99) within the Switzerland to USA transfer, 82.83 % (95% CI 81.97 to 83.67) within the Switzerland to France transfer, 85.73 % (95% CI 85.08 to 86.34) within the France to USA transfer and 82.28 % (95% CI 80.04 to 84.15) within the France to Switzerland transfer. All validation sample accuracies with mean accuracies and CIs are shown in figure 2 stratified by transfer combination.



**Figure 2: Validation sample accuracies stratified by transfer combination.** Each black dot represents the validation sample accuracy for one transfer with the red dot representing the mean accuracy. The bars indicate the 95% CI around the validation sample accuracies.

## Predicted validation sample accuracies

We found that our method’s predicted validation sample accuracy mean to be 83.22 % (95% CI 81.1 to 85.39) within the USA to Switzerland transfer, 81.8 % (95% CI 80.41 to 83.47) within the USA to France transfer, 99.38 % (95% CI 99.02 to 99.78) within the Switzerland to USA transfer, 99.57 % (95% CI 99.22 to 99.86) within the Switzerland to France transfer, 88.21 % (95% CI 87.19 to 89.23) within the France to USA transfer and 91.19 % (95% CI 90 to 92.49) within the France to Switzerland transfer. All predicted validation sample accuracies with mean accuracies and CIs are shown in figure 3 stratified by transfer combination.

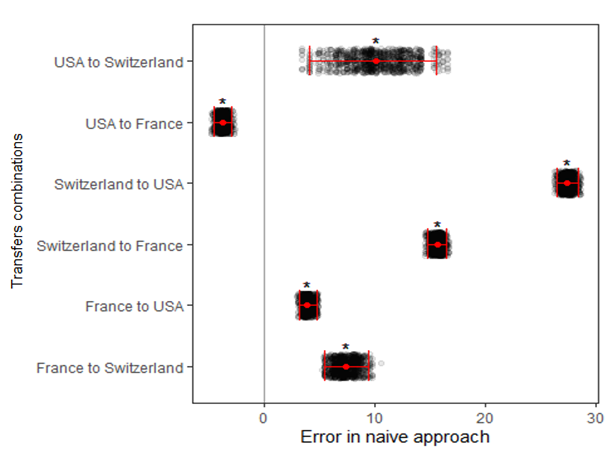


**Figure 3: Predicted validation sample accuracies stratified by transfer combination.** Each black dot represents our method’s predicted validation sample accuracy in one transfer with the red dot representing the mean accuracy. The bars indicate the 95% CI around our method’s predicted validation sample accuracies.

## Error in naive approach

We found that the development sample accuracies significantly overestimated the validation sample accuracies in five of our six transfer combinations. The mean errors in these five transfers combinations were 10.19 (95% CI 4.15 to 15.63) within the USA to Switzerland transfer, 27.43 (95% CI 26.5 to 28.41) within the Switzerland to USA transfer, 15.66 (95% CI 14.81 to 16.47) within the Switzerland to France transfer, 3.95 % (95% CI 3.18 to 4.81) within the France to USA transfer and 7.4 (95% CI 5.54 to 9.43) within the France to Switzerland transfer.

We found that the development sample accuracies significantly underestimated the validation sample accuracies in the remaining transfer combination. The mean error in this transfer combination was -3.69 (95% CI -4.45 to -2.92) within the USA to France transfer. All errors in the naive approach with mean errors and CIs are shown in figure 4 stratified by transfer combination.

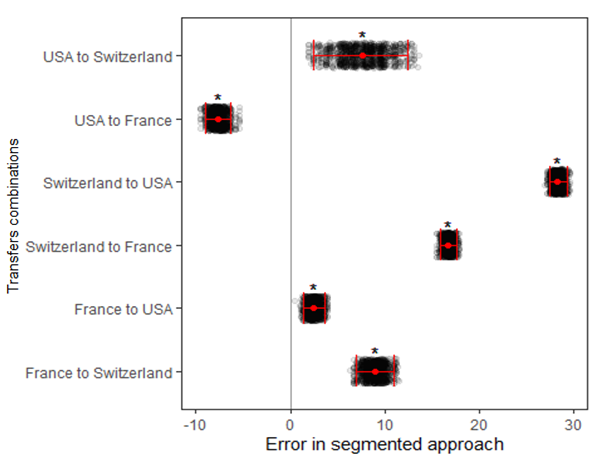


**Figure 4: Errors in the naive approach stratified by transfer combination.** Each black dot represents the error between the development sample accuracy and the validation sample accuracy for one transfer with the red dot representing the mean error across all transfers. The bars indicate the 95% CI around the errors. Asterisk (\*) above the mean error indicate statistical significance.

## Error in segmented approach

We found that our method’s predicted validation sample accuracies significantly overestimated the validation sample accuracies in five of our six transfer combinations. The mean errors in these five transfer combinations were 7.63 (95% CI 2.43 to 12.46) within the USA to Switzerland transfer, 28.32 (95% CI 27.42 to 29.31) within the Switzerland to USA transfer, 16.74 (95% CI 15.89 to 17.65) within the Switzerland to France transfer, 2.48 (95% CI 1.37 to 3.66) within the France to USA transfer and 8.91 (95% CI 7.04 to 10.96) within the France to Switzerland transfer.

We found that our method’s predicted validation sample accuracies significantly underestimated the validation sample accuracies in the remaining transfer. The mean error in this transfer combinations were -7.66 (95% CI -8.97 to -6.26) within the USA to France transfer All errors in the naive approach with mean errors and CIs are shown in figure 5 stratified by transfer combination.

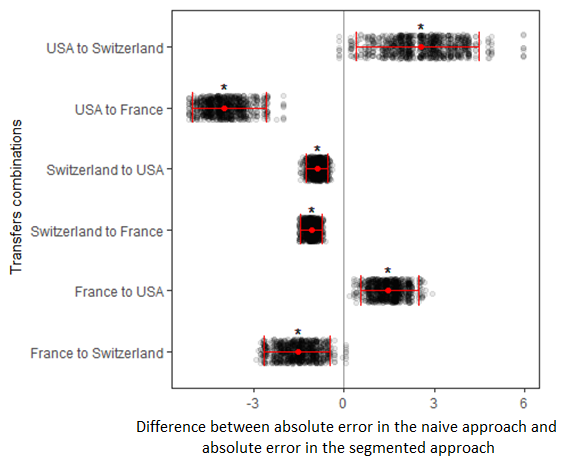


**Figure 5: Errors in the segmented approach stratified by transfer combination.** Each black dot represents the error between our method’s predicted validation sample accuracy and the validation sample accuracy for one transfer with the red dot representing the mean error across all transfers. The bars indicate the 95% CI around the errors. Asterisk (\*) above the mean error indicate statistical significance.

## Difference between absolute error in the naive approach and absolute error in the segmented approach

We found that our segmented approach was significantly better than the naive approach in two of our six transfer combinations. The mean differences in error in these transfers were 2.56 (95% CI 0.39 to 4.5) in the USA to Switzerland transfer and 1.47 (95% CI 0.54 to 2.47) in the France to USA transfer.

We found that our segmented approach was significantly worse than the naive approach in four of our six transfer combinations. The mean difference in error in these transfers were -3.97 (95% CI -5.02 to -2.57) in the USA to France transfer, -0.89 (95% CI -1.27 to -0.54) in the Switzerland to USA transfer, -1.09 (95% CI -1.46 to -0.72) in the Switzerland to France transfer and -1.51 (95% CI -2.67 to -0.47) in the Switzerland to USA transfer. All error differences with mean error differences and CIs are shown in figure 6 stratified by transfer combination.



**Figure 6: Difference between the absolute error in naive approach and the absolute error in the segmented approach stratified by transfer combination.** Each black dot represents the difference between the absolute naive approach and the absolute segmented approach for one transfer with the red dot representing the mean difference across all transfers. The bars indicate the 95% CI around the differences. Asterisk (\*) above the mean difference indicate statistical significance.

# Discussion

## Key findings

At present, no methods exist that can predict the performance of a prediction model after transfer using unlabeled data. In this study, we have developed and tested such a method. We found that our method’s predicted accuracy was significantly worse at predicting the accuracy after transfer compared to the development sample accuracy. This was the case four of our six transfer combinations. These results suggest against our hypothesis that our method’s predicted accuracy would be as good as or better than the development sample accuracy at predicting accuracy after transfer.

## Naive approach and segmented approach

In the naive approach, we found that the development sample accuracies significantly overestiamted the validation sample accuracies in most of our transfer combinations. These results suggest that our initial predictions were mostly optmistic, which corroborates with the concluding remarks of Moons et al (13), that most prediction models perform worse when they are applied to new individuals that differ from those used to train the prediction model. As an example, Ohnuma et al (22) showed similar losses in performance in most of the prediction models that had been externally validated in their review.

In the segmented approach, we found that our method’s predicted accuracies significnatly overestimated the validation sample accuracies in most of our transfer combinations. These results suggest that our method’s predictions were still mostly optimsitc. One reason why our method’s predictions were still optimistic may have been because we included participants from the segmented samples in the training of the prediction models. If the prediction models were overfitted to these participants, this may have been the reason why our predictions were still optmistic (12). Another reason may have been that the association between the model predictors and model outcomes varied greatly between the segmented samples and the validation samples. This could be due to differences in the healthcare systems, how the model predictors were measured, patient characteristics and problems that may have arised during the sample simulations. This together with the fact that the prediction models were trained to learn the association better within the segmented samples may have been the reason why our method’s predictions were still optimistic.

## Transfer combinations where our method was better

Because most prediction models perform worse when they are applied to new individuals, one could argue that it would be beneficial for our method to predict an accuracy that is lower than the development sample accuracy in most transfers. In the two transfer combinations where our method’s predicted accuracy was significantly better than the development sample accuracy, this was the case.

One reason to why our method’s predicted accuracy was lower in these transfer combinations, may have been because a smaller proportion of participants from the segmented samples were included in the training of the prediction models. Another reason may have been that the association between the model predictors and the model outcomes were more similar between the segmented samples and the validation samples in these transfer combinations. Another reason may have been that our propensity models misclassified fewer validation samples than development samples in these transfer combinations. Because the misclassified validation samples have more similar distribution of model predictors as the development samples that were not included in the segmented samples, having fewer of these may have made the segmented samples better representatives of the validation samples.

## Methods predicting performance

There are several methods that propose to predict the accuracy of a prediction model more accurately in new individuals than the accuracy within the individuals that were used to train the prediction model. These methods include different types of split-sampling, cross-validation and boostrapping (23-26). However, these methods only predict accuracy more accurately in individuals that are similar to those that were used to train the prediction model, and not after the prediction model have been transferred to a different setting (27).

If these methods were to be used, especially cross-validation, to predict the accuracy after transfer, one would expect accuracies that are similar to our development sample accuracies. This is because of our development sample accuracies being derived somewhat similarly.

## Prediction model selection strategy

Although not being a usual way of selecting a prediction model, the way that we selected prediction models based on cross-validation should have resulted in less optimistic accuracies in the development samples. This is because we used fewer participants from the development samples to train the prediction models than what was used to assess the accuracy with within the development samples. This should have made it either as difficult or more difficult for our method to predict an accuracy that is as good as or better than the development sample accuracy at predicting the accuracy after transfer.

Even if we did not succeeded in selecting prediction models that were less optimistic in the development samples, it should have not mattered as our method’s predicted accuracies were still significantly worse than the development sample accuracies in most of the transfer combinations.

## Strengths and limitations

The strength of our study is that to the best of our knowledge, this is the first study to develop and test a method that predicts the predictive performance of a prediction model after transfer using unlabeled data. Such a method could provide information that can be used to simplify the implementation of prediction models in health care and therefore indirectly improve decision making changes in health care professionals, patient health outcome or cost-effectiveness of care, if the impact of the model is sufficient.

Our study also has limitations. First, because our original sample sizes were small in some of the countries, we chose to simulate new samples. These simulate samples were probably poor representatives of real samples as we simulated outcomes using logistic regression models trained with the original samples. Logistic regression models that are trained with few events per variable, which the original small samples had, has frequently been associated with poorer predictive performance (28). The simulated outcomes in the countries that originally had small sample sizes may therefore have been of poor quality, which may have reflected poorly on our method. Second, based on simplicity, we chose to only include model predictors that were continuous and that had no missing data. Although there is no consensus on which method is best for selecting model predictors, it has been recommended to include all available model predictors to reduce overfitting and selection bias (12). Because we did not include categorical variables, we may have increased overfitting and selection bias which may have reflected poorly on our method. Third, because we did not assess the proportion of participants from the segmented samples that were used to train the prediction models, we can only hypothesize how this may have reflected on our results. Fourth, because we did not assess the number of development samples and validation samples that were misclassified by our propensity model, we may have identified segmented samples that did not have more similar distributions of model predictors as in the validation samples. Fifth, due to simplicity, we chose to only assess the performance of our prediction models in terms of the number of correct predictions compared to the total number of predictions for just one decision threshold. Last, due to being limited by time, we decided to only test our method in one dataset with data from three different countries.

## Significance

Because we have found results that contradict our hypothesis, we cannot recommend that our method is used to predict the accuracy of a prediction model after transfer.

## Future studies

Because our method may have been negatively affected by several factors in this study, a future study may reduce these factors to gain a better understanding of the true performance of our method. These factors may be reduced by using a dataset with larger samples, by including all model predictors that are available in the dataset and by testing the method for several prediction models. Further improvements could also be done by assessing the performance in terms that are independent of the decision threshold, such as the area under the receiver operating characteristic curve.

Because we did not assess how the proportion of participants from the segmented samples that were used to train the prediction model may have affected our method, a future study could separate transfers based on this proportion to gain a better understanding of the performance of our method and possible improvements to it. Similarly, because we did not assess the relationship between the number of development samples and validation samples that were misclassified by our propensity model, a future study could separate transfers based on this relationship to gain a better understanding of the performance of our method and possible improvements to it.

# Conclusions

Although our method used unlabeled data to predict the accuracy of a prediction model after transfer, our results do not support that our method is used for such predictions. However, because of the limitations of this study that may have negatively affected our method, further studies are needed to gain a better understanding of our method’s true performance and possible areas of improvement.

# Contributions

All statistical analysis except the simulation of new samples were done by Arman Norouzi. The sample simulations were done by Martin Gerdin Wärnberg.

# Acknowledgements

I would like to thank my supervisor Martin Gerdin Wärnberg for the opportunity to take part in this project. He has been available for questions and discussion almost daily while giving feedback weekly. My thanks are extended to Andreea Bocancea for giving feedback. I would also like to thank my coordinator Mark Taylor for his valuable feedback.

# References

1. Steyerberg EW, Moons KG, Windt DA van der, Hayden JA, Perel P, Schroter S, et al. Prognosis research strategy (progress) 3: Prognostic model research. PLoS Med. 2013;10(2):e1001381.

2. Moons KG, Royston P, Vergouwe Y, Grobbee DE, Altman DG. Prognosis and prognostic research: What, why, and how? Bmj. 2009;338:b375.

3. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (tripod) the tripod statement. Circulation. 2015;131(2):211–9.

4. Deo RC. Machine learning in medicine. Circulation. 2015;132(20):1920–30.

5. Riley RD, Hayden JA, Steyerberg EW, Moons KG, Abrams K, Kyzas PA, et al. Prognosis research strategy (progress) 2: Prognostic factor research. PLoS Med. 2013;10(2):e1001380.

6. Shell IG, Greenberg GH, McKnight RD, Nair RC, McDowell I, Reardon M, et al. Decision rules for the use of radiography in acute ankle injuries: Refinement and prospective validation. Jama. 1993;269(9):1127–32.

7. Steyerberg E. A practical approach to development, validation, and updating. New York: Springer; 2009.

8. Dorresteijn JA, Visseren FL, Ridker PM, Wassink AM, Paynter NP, Steyerberg EW, et al. Estimating treatment effects for individual patients based on the results of randomised clinical trials. Bmj. 2011;343:d5888.

9. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC medical research methodology. 2006;6(1):18.

10. 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.

11. Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, et al. 2016 esc guidelines for the management of atrial fibrillation developed in collaboration with eacts. European journal of cardio-thoracic surgery. 2016;50(5):e1–e88.

12. Royston P, Moons KG, Altman DG, Vergouwe Y. Prognosis and prognostic research: Developing a prognostic model. Bmj. 2009;338.

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;98(9):691–8.

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. Journal of clinical epidemiology. 2003;56(5):441–7.

15. Steyerberg E, Eijkemans M, Habbema J. Application of shrinkage techniques in logistic regression analysis: A case study. Statistica Neerlandica. 2001;55(1):76–88.

16. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. American heart journal. 1991;121(1):293–8.

17. Eckart A, Hauser SI, Kutz A, Haubitz S, Hausfater P, Amin D, et al. Combination of the national early warning score (news) and inflammatory biomarkers for early risk stratification in emergency department patients: Results of a multinational, observational study. BMJ open. 2019;9(1):e024636.

18. Dryad [internet]. Data from: Combination of the national early warning score (news) and inflammatory biomarkers for early risk stratification in emergency department patients: Results of a multi-national, observational study. 2018 [cited 2020 Dec 10]; Available from: <https://datadryad.org/stash/dataset/doi:10.5061/dryad.d22q6vh>

19. R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2020. Available from: <https://www.R-project.org/>

20. Venables WN, Ripley BD. Modern applied statistics with s [Internet]. Fourth. New York: Springer; 2002. Available from: <http://www.stats.ox.ac.uk/pub/MASS4/>

21. Kuhn M. Caret: Classification and regression training [Internet]. 2020. Available from: <https://CRAN.R-project.org/package=caret>

22. Ohnuma T, Uchino S. Prediction models and their external validation studies for mortality of patients with acute kidney injury: A systematic review. PLoS One. 2017;12(1):e0169341.

23. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American statistical association. 1983;78(382):316–31.

24. Efron B, Tibshirani RJ. An introduction to the bootstrap. CRC press; 1994.

25. Efron B, Tibshirani R. Improvements on cross-validation: The 632+ bootstrap method. Journal of the American Statistical Association. 1997;92(438):548–60.

26. Picard RR, Berk KN. Data splitting. The American Statistician. 1990;44(2):140–7.

27. Altman DG, Vergouwe Y, Royston P, Moons KG. Prognosis and prognostic research: Validating a prognostic model. Bmj. 2009;338:b605.

28. Smeden M van, Moons KG, Groot JA de, Collins GS, Altman DG, Eijkemans MJ, et al. Sample size for binary logistic prediction models: Beyond events per variable criteria. Statistical methods in medical research. 2019;28(8):2455–74.