Supervisor: Martin Gerdin Wärnberg

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

End 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:* För att implementera prediktionsmodeller inom sjukvården, måste prediktionsmodellens prestanda bedömas i ett nytt kontext. Denna prestanda efter överföring skulle vara önskvärd att förutsäga eftersom att det skulle kunna förenkla implementeringen av prediktionsmodell inom sjukvården. Den nuvarande methoden som använder en naive tillvägagångsätt är optimistisk och använder inte omärkt data vilket väckt frågan om det är möjligt att utveckla en metod som är mer exakt, genom att använda omärkt data. *Syfte*: Att utveckla och testa en ny metod för att förutsäga prestanda av en prediktionsmodell efter överföring till ett annat land med omärkt data. *Material och Metoder:* Vi använde oss av ett data med deltagare från tre länder för att simulera ny data i varje land. Denna data användes för att utveckla prediktionsmodeller i varje land för att sedan bedöma prestandan av modellerna inom det landet som de utvecklat i, i det landet som de överförts till samt i ett segment av det land som de utvecklats i. Dessa prestandor jämfördes och hela processen upprepades 1000 gånger.

*Resultat*: Vi fann att vår metod, som använde sig av ett segmenterat tillvägagångsätt, var signifikant bättre än det naiva tillvägagångsättet på att förutsäga prestanda efter överföring i två av våra sex överföringskominationer. I de återstående fyra överföringskombinationerna fann vi resultat som var exakt tvärtom. *Slutsats*: Resultaten i denna studie stödjer inte anvädningen av vår metods segmenterade tillvägagångsätt för att förutsäga prestandan av en prediktionsmodell efter överföring till ett nytt kontext. Dock krävs ytterligare studier för att få en bättre förståelse för vår metod.

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

*Introduction:* To implement prediction models in health care, the prediction models performance must be assessed in a new setting. This performance after transfer would be desirable to predict as it could simplify the process of implementing prediction models in health care. The current method that uses a naive approach is optimistic and does not use unlabelled data. This raised the question if it is possible to develop a method that is more accurate by using unlabelled data. *Aims:* To develop and test a new method that predicts prediction model performance after transfer using unlabelled data. *Material and Methods:* We used a dataset with samples from three countries to simulate new samples in each country. These samples were used to develop prediction models, and each model’s performance was assessed in the country that it was developed in, in the country it was transferred to, and in a segment of the country that it was developed in. These performances were compared, and the processes was repeated 1000 times. *Results:* We found in two of our six transfer combinations that our methods segmented approach was significantly better than the naive approach at predicting performance after transfer. In the remaining four transfer combinations, we found results that were the exact opposite. *Conclusions:* The results in this study do not support the use of our method for predicting prediction model performance after transfer. However, due to the limitations of this study, further studies are required to gain a better understanding of our method.

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

# Abbreviations

CHA2DS2-VASc congestive heart failure, hypertension, age >74 (2 points), diabetes, stroke/transiet ischemi attack/thromboembolism (2 points), vascular disease, age 65-74, sex (female)

CI Confidence interval

ED Emergency department

# 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 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). Prediction models in health care can be defined as statistical algorithms that combine multiple predictors (symptoms and characteristics) by assigning relative weights to each predictor in order to predict the risk of a specific outcome in a patient (1-3). The prediction models are capable of this task due to being trained to find the association between the predictors and the outcome of interest. These associations can then be used to predict outcome in new patients based on the new patients unlabelled predictors (4).

### Diagnostic and prognostic models

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

In a prognostic model the predicted risk can be used to choose between therapeutic options, make lifestyle changes and to risk-stratify patients in therapeutic clinical trials (2, 6-8). An example of a prognostic model is the CHA2DS2-VASc score. This prediction model help predict the annual risk for ischemic stroke in patients with atrial fibrillation. To predict this risk the prediction model uses predictors such as history for congestive heart failure, hypertension, age >74, diabetes, stroke/transient ischemic attack/thromboembolism, vascular disease, age 65-74 and female sex (9). Based on the predicted risk, health care professionals can decide whether the patient needs anticoagulation treatment (10).

## Prediction models studies

There are many uses for prediction models in the field of medicine, where the Ottawa Ankle Rules and the CHA2DS2-VASc score are just two examples of prediction models that have been implemented in clinical practice. To develop and implement such useful models within health care, several studies are needed to be carried out. These studies involve prediction model development studies, validation studies and impact studies (11, 12).

### Prediction model development study

The first study that is needed to be carried out is a prediction model development study. In these studies, the aim is to develop a prediction model by training a statistical algorithm with a development sample that consist of several patients with registered predictors labelled with an outcome of interest (11). To identify these relevant predictors there are two main strategies: one that is clinically driven and another that is data driven. In the clinically driven strategy, the predictors are identified and selected by clinical experts in a research group or by literature review. In the data driven strategy, all predictors are initially included (13).

After the development sample has been defined, the statistical algorithm must be chosen. There is no consensus on which algorithm one should chose, but generally when the data is limited in the development sample a simpler algorithm is used such as a logistic regression (4). When the algorithm has been trained with the development sample and the model has been developed, it is generally optimistic in its predictive performance within the development sample when compared with the performance in a sample that is independent from the development sample (14). This optimism is a result of overfitting, which means that the model is too closely adapted to sample used to train the model (11). It is therefore important to quantify this optimism through internal validation techniques, in order to get a better prediction of the performance of the model in similar patients (6). This quantified optimism can thereafter be adjusted for in the development study by applying shrinkage or penalization to the prediction model (15).

### Prediction model validation study

The second study that is needed to be carried out is a prediction model validation study. In these studies, the aim is to assess the generalizability or transportability of the prediction model. To assess the generalizability or transportability of the prediction model, the model must be transferred to a new hospital or country in which the performance is assessed in a validation sample. This sample consist of new patients with registered predictors labelled with the outcome of interest. This type of external validation is important to perform as most prediction models perform worse when they are applied to new patients that are different from those that were used to train the prediction model (12).

### Prediction model impact study

The last study that is needed to be carried out is a prediction model impact study. In these studies, the aim is to assess the impact of the model in clinical practice, ideally in a randomized trial. The impact of the model is assessed in measures such as improvements in patient health outcomes and/or cost-effectiveness of care. These studies are carried out to prove that the model is of value in clinical practice (12).

## Problems with 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 arise during the prediction model validation study. In these studies, both predictors and the outcome that the predictors have been labelled with is required from the transfer setting in order to assess the generalizability or transportability of the prediction model (12). This data is not always available retrospectively and can present a problem that is time inefficient and expensive if the data is difficult to access when collecting it prospectively.

This would for example be a problem if the data for the Framingham Risk Score were to be collected prospectively in order to assess the generalizability or transportability of this model. The predictors in this model are cheap blood samples and simple demographics while the outcome data is cardiovascular disease within 10 years (16). Although having predictors that are easily accessible, the outcome is not, as it requires 10 years of follow up to be collected.

Because the outcome data can be difficult to access when the data is collected prospectively, it would be desirable to have a method that could predict the performance of the prediction model after transfer without the need for outcome data from the transfer setting. Such a method could simplify the process of implementing prediction models in health care and therefore indirectly improve patient health outcomes and cost-effectiveness of care, if the impact of the model is sufficient.

The only method that is currently available for predicting prediction model performance after transfer without outcome data, is a method that uses a “naive approach”. This naive approach assumes that the performance after transfer is the same as the performance that can be assessed in the development sample. This method is optimistic, as it has been shown that most prediction models perform worse when they are applied to new patients that are different from those that were used to train the prediction model (12).

This raised the question if it is possible to develop a method that uses unlabelled data (predictors that have not been labelled with outcome) from the transfer setting, in order to predict the performance after transfer more accurately when compared with the naive approach. At present, no such method exist 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 unlabelled data.

### Hypothesis

We hypothesized that our new method, that uses a “segmented approach” to predict prediction model performance after transfer with unlabelled data, would be as good as or better than the naive approach. This hypothesis was because the segmented approach would predict a performance that is derived from a segment identified in the development sample, instead of the complete development sample, that consist of observations that are more similar to the observations from the transfer setting. Because more similar observations should lead to more similar predictions, we hypothesized that predictions in this segment would result in a performance that is more similar to the performance after transfer.

# Aim

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

# Methods and Materials

## Study design

The study design was a simulation study. The analysis in this study was performed using a dataset from a multinational observational study that has been made freely reusable by Eckert A et al in the Dryad Digital Repository (17, 18). This dataset was chosen due to two reason: the first reason being that it has been previously used to develop prediction models, and the second reason being that it contains patient data from three different countries.

## Participants

The patients enrolled in the dataset that we used 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 predictors that were used from the dataset to develop prediction models and simulate new 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. These predictors were all measured during the patient’s time in the ED. How they were measured was not mentioned in the study that publicized the dataset.

These specific predictors were selected because of two reasons: the first being that they were all the available predictors that were continuous in the dataset and the second being that they had no missing data. Although no consensus exits for which method one should chose when selecting predictors for a prediction model, it has been recommended to include all available predictors to reduce the model from adjusting too closely to the sample that is used for training (11). To follow this recommendation while still simplifying the analysis, the categorical predictors were excluded.

### Model outcomes

The outcome that was used from the dataset to develop prediction models and simulate new outcome in the statistical analysis was ICU admission. This specific outcome was selected from the dataset because of two reasons: the first being that it is was a binary outcome which was a requirement for our prediction models and the second being that it was the most frequent occurring binary outcome available in the dataset. However, any of the other binary outcomes available in the dataset could have been selected.

To develop the propensity model in the statistical analysis, the country in which the patient sought ED care was used. This specific outcome was selected for the propensity model because of two reasons: the first being that it could be collected without being expensive and second being that it could be used to predict form which country the patient sought ED care.

### Missing data

Because of the dataset was 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 patients sought ED care (USA sample, France sample and Switzerland sample).

### Decision threshold

The decision threshold was set to 0.5 for the prediction models and the propensity models that were developed during the statistical analysis. This decision threshold dichotomizes the predictions made by the models based on the score calculated with the predictors. This threshold could be anything between 0 to 1. This specific decision threshold was selected to simplify the analysis.

### Measurement of performance

The performance of the prediction models was measured in accuracy (number of correct predictions divided by total number of predictions).

### Sequence of analysis

Analysis in this study was performed in the programming language R (19). The sequence of analysis performed were sample simulation, sample assignment, prediction model development, development sample performance, validation sample performance, propensity model development, predicted validation sample performance 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 predictor simulation and a outcome simulation. The 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 outcome that was used to simulate new model outcomes were ICU admission from the divided samples.

To perform the 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 predictors in the divided sample to simulate new predictors. To perform the outcome simulation for the simulated predictors, the glm function implemented in R was used to develop a logistic regression model by training the model with the divided sample that was used to simulate the predictors. This model was used to predict outcome based on the simulated predictors. The predictions made were used to label the predictors with an outcome. The simulated predictors labelled with the simulated outcome constituted as a simulated sample for the divided sample used to simulate both predictors and outcome. This process of simulation was 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 developed, 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 model from adapting to closely to the development sample, cross-validation was used 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 in the country that it was developed in, the model developed in the prediction model development step was used to predict outcome based on predictors in the development sample. These predictions were compared with the true outcome 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 based on predictors in the validation sample. These predictions were compared with the true outcome 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. The pooled sample was used to develop a propensity model, also a prediction model, by training a logistic regression model with the pooled sample. The propensity model was then used to predict the origin of the samples in the pooled sample. Observations from the development sample that were misclassified as validation sample observations, were used to identify a segmented sample. This segmented sample consists of observations in the development sample that are more similar to the observations in 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.

### Segmented sample accuracy

To assess our methods prediction of the accuracy in the validation sample, the model developed in the prediction model development step was used to predict outcome based on predictors in the segmented sample. These predictions were compared with the true outcome in the segmented sample in order to acquire our methods prediction of the accuracy in the validation sample.

### Naive approach

To assess how well the naive approach predicts the performance after transfer, the difference between the accuracy in the development sample and the accuracy in the validation sample was calculated.

### Segmented approach

To assess how well the segmented approach predicts the performance after transfer, the difference between the accuracy in the segmented sample and the accuracy in the validation sample was calculated.

### Approach comparison

To assess the difference in performance between the two approaches, 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 unlabelled data. Such a method could simplify the process of implementing prediction models in health care and therefore indirectly improve patient health outcomes and/or cost-effectiveness of care if the impact of the model is sufficient.

### 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

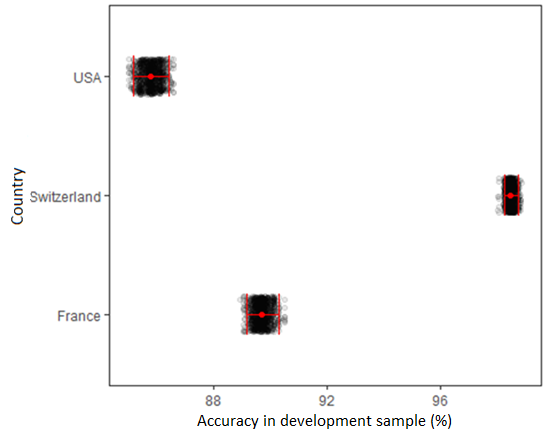
With the 1303 participants in the original dataset, 10 000 000 new participants were simulated for each country. The mean age of the participants in the simulated sample was 63.84 years. Baseline characteristics of the simulated samples stratified by country are shown in table 2.

**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 accuracy

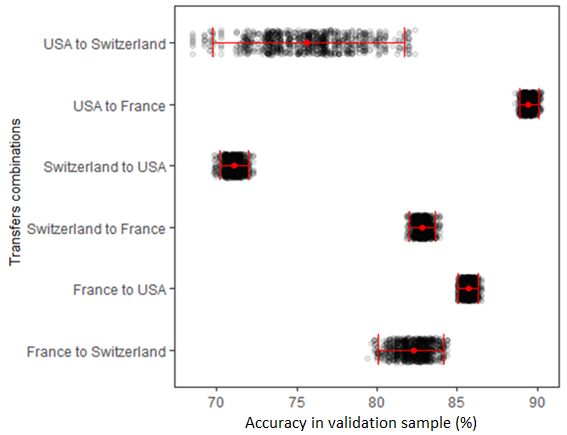
We found the mean development sample accuracy to be 85.78% (95% CI 85.17 to 86.39) within USA, 98.48% (95% CI 98.26 to 98.76) within Switzerland and 89.68% (95% CI 89.16 to 90.27) within France. This accuracy is the accuracy that the naive approach uses as a prediction of the accuracy after transfer. The mean accuracy in the development sample with CIs around the accuracies are shown in figure 1 stratified by country.



**Figure 1: Development sample accuracy stratified by country.** The black dots represent the accuracy in the development sample for one sequence of the analysis. The red dot represents the mean development sample accuracy. The bars indicate the 95% CI around the accuracies.

## Validation sample accuracy

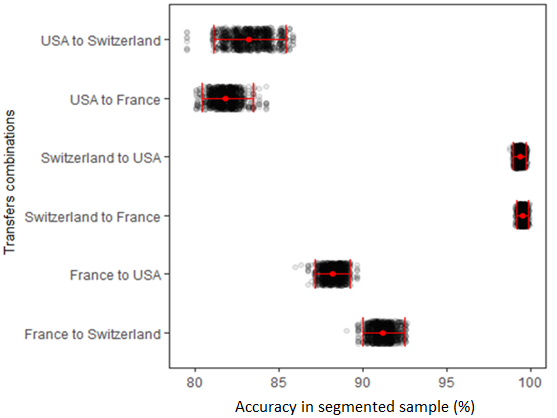
We found that the mean validation sample accuracy to be 75.59% (95% CI 69.72 to 81.71) within the USA to Switzerland transfer, 89.46% (95% CI 88.91 to 90.08) within the USA to France transfer. 71.05% (95% CI 70.19 to 71.99) within the Switzerland to USA transfer, 82.83% (95% CI 81.97 to 83.67) within 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. This accuracy is the accuracy that both the naive approach and the segmented approach is trying to predict. The mean validation sample accuracy with CIs around the accuracies are shown in figure 2 stratified by transfer combination.



**Figure 2: Validation sample accuracy stratified by transfer combination.** The black dots represent the validation sample accuracy for one sequence of the analysis. The red dot represents the mean validation sample accuracy. The bars indicate the 95% CI around the accuracies.

## Segmented sample accuracy

We found that the mean segmented sample accuracy 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. This accuracy is the accuracy that the segmented approach uses as a prediction of the accuracy after transfer. The mean segmented sample accuracy with CIs around the accuracies are shown in figure 3 stratified by transfer combination.

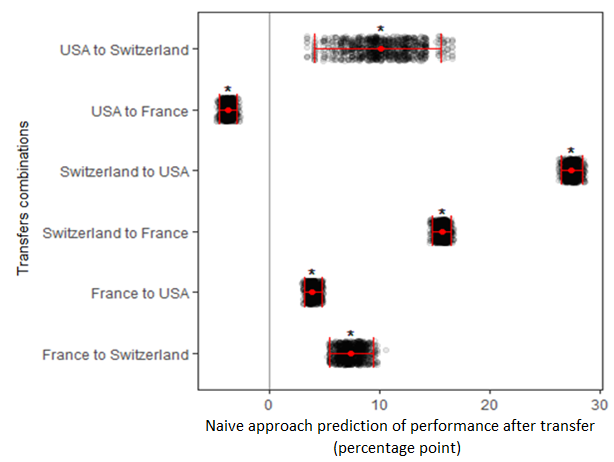


**Figure 3: Segmented sample accuracy stratified by transfer combination.** The black dots represent the segmented sample accuracy for one sequence of the analysis. The red dot represents the mean segmented sample accuracy. The bars indicate 95% CI around the accuracies.

## Naive approach

We found that the naive approach, that assumes that the performance after transfer is the same as the performance in the development sample, significantly overestimated the performance after transfer in five of our six transfer combinations. In these transfer combinations, we found that the naive approach overestimated the performance after transfer with a mean of 10.19 percentage points (95% CI 4.15 to 15.63) within the USA to Switzerland transfer, 27.43 percentage points (95% CI 26.5 to 28.41) within the Switzerland to USA transfer, 15.66 percentage points (95% CI 14.81 to 16.47) within the Switzerland to France transfer, 3.95 percentage points (95% CI 3.18 to 4.81) within the France to USA transfer and 7.4 percentage points (95% CI 5.54 to 9.43) within the France to Switzerland transfer.

We also found that the naive approach significantly underestimated the performance after transfer in one of our six transfer combinations. In this transfer combination, we found that the naive approach underestimated the performance after transfer with a mean of -3.69 percentage points (95% CI -4.45 to -2.92) within the USA to France transfer. The naive approach prediction of the performance after transfer is shown in figure 4 stratified by transfer combination.

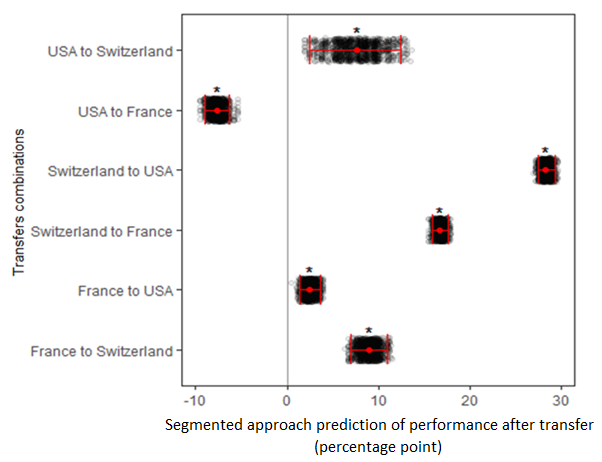


**Figure 4: Naive approach prediction of the performance after transfer stratified by transfer combination.** The black dot represents the difference between the development sample accuracy and the validation sample accuracy for one sequence of the analysis. The red dot represents the mean percentage point difference. The bars indicate 95% CI around the percentage point differences. Asterisk (\*) above the mean percentage point difference indicate statistical significance.

## Segmented approach

We found that our methods segmented approach, that uses the performance in the segment as a prediction of the performance after transfer, significantly overestimated the performance after transfer in five of our six transfer combinations. In these transfer combinations, we found that the segmented approach overestimated the performance after transfer with a mean of 7.63 percentage points (95% CI 2.43 to 12.46) within the USA to Switzerland transfer, 28.32 percentage points (95% CI 27.42 to 29.31) within the Switzerland to USA transfer, 16.74 percentage points (95% CI 15.89 to 17.65) within the Switzerland to France transfer, 2.48 percentage points (95% CI 1.37 to 3.66) within the France to USA transfer and 8.91 percentage points (95% CI 7.04 to 10.96) within the France to Switzerland transfer.

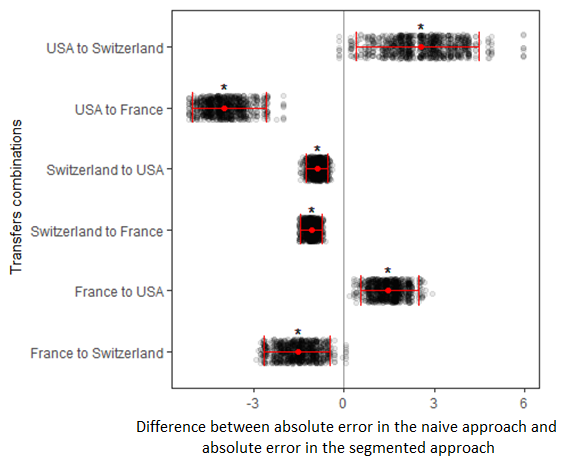
We also found that the segmented approach significantly underestimated the performance after transfer in one of our six transfer combinations. In this transfer combination, we found that the segmented approach underestimated the performance after transfer with a mean of -7.66 (95% CI -8.97 to -6.26) within the USA to France transfer. The segmented approach prediction of the performance after transfer is shown in figure 4 stratified by transfer combination.

**Figure 5: Segmented approach prediction of the performance after transfer stratified by transfer combination.** The black dot represents the difference between the segmented sample accuracy and the validation sample accuracy in percentage points for one sequence of the analysis. The red dot represents the mean difference. The bars indicate 95% CI around the percentage point differences. Asterisk (\*) above the mean percentage point difference indicate statistical significance.

## Approach comparison

We found that our segmented approach was significantly better than the naive approach at predicting the performance after transfer in two of our six transfer combinations. In these transfers combinations, the mean better prediction of the performance after transfer by the segmented approach were 2.56 percentage points (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 also found that our segmented approach was significantly worse than the naive approach at predicting the performance after transfer in four of our six transfer combinations. In these transfer combinations, the mean worse prediction of the performance after transfer by the segmented approach were -3.97 percentage points (95% CI -5.02 to -2.57) in the USA to France transfer, -0.89 percentage points (95% CI -1.27 to -0.54) in the Switzerland to USA transfer, -1.09 percentage points (95% CI -1.46 to -0.72) in the Switzerland to France transfer and -1.51 percentage points (95% CI -2.67 to -0.47) in the Switzerland to USA transfer. The difference between the naive approach and the segmented approach prediction of the performance after transfer are shown in figure 6 stratified by transfer combination.

**Figure 6: Difference between the naive approach and the segmented approach prediction of the performance after transfer.** The black dot represents the difference between the absolute value of the naive approach and the absolute value of the segmented approach for one sequence of the analysis. The red dot represents the mean difference.The bars indicate 95% CI around the differences. Asterisk (\*) above mean difference indicate statistical significance.

# Discussion

## Key findings

To implement prediction models in health care, it is required to assess the performance of the prediction model in new patients after the model has been transferred to a new hospital or country (12). This performance after transfer can be time inefficient and expensive to assess if the outcome data is difficult to access from the new patients. It would therefore be desirable to have a method that could predict the performance of a prediction model after transfer without the use of outcome data, as it could simplify the process of implementing prediction models in health care. Currently, the only method capable of such predictions is a method that uses a naive approach. This method is optimistic and does not use unlabelled data. This raised the question if it is possible to develop a better method, that uses unlabelled data in order to make better predictions of the performance after transfer. In this study, we have developed and tested such method.

We found that our method segmented approach that uses unlabelled data, were significantly worse than the naive approach at predicting the performance of a prediction model after transfer. This was the case in four of our six transfer combinations. These results suggest against our hypothesis, that our segmented approach would be as good as or better than the naive approach at predicting the performance after transfer.

## Naive approach

In the naive approach, we found that most of our initial prediction of the performance after transfer were significantly overestimating the true performance after transfer. These findings suggest that our initial predictions were mostly optimistic, which corroborates with the findings of Moons et al (12), that most prediction models perform worse when they are applied to new patients that are different from those that were used to develop 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.

## Segmented approach

In the segmented approach, we found that most of our methods prediction of the performance after transfer were significantly overestimating the true performance after transfer. These findings suggest that most of our methods prediction of the performance after transfer were optimistic, even though unlabelled data was used to identify a segment that would have observations more similar to the observations after transfer. One possible explanation to why the segmented approach was still mostly optimistic, may have been because we included patients from the segmented samples in the training of the prediction models. If the models were too closely adapted to these patients, it may have been the reason to why our new methods predictions were still mostly optimistic.

Another possible explanation to why the segmented approach was still mostly optimistic, may have been because the association between the predictors and the outcome varied greatly between the countries. If this association varied greatly between the countries, it would have mattered less if we succeeded in identifying a segment of observations that were more similar to the observations after transfer, because the outcome would have not been more similar.

## Transfer combinations where the segmented approach was better

Because most prediction models perform worse when they are applied to new patients, one could argue that it would be beneficial for our methods segmented approach to predict a performance that is less optimistic in most transfers. In the two transfer combinations where the segmented approach was significantly better than the naive approach at predicting the performance after transfer, this was the case.

One possible explanation to why our methods segmented approach was less optimistic in these transfer combinations, may have been because fewer patients from the segmented samples being included in the training of the prediction model. This may have reduced the effect of the model adapting to closely to the training data and therefore why the predicted performance was less optimistic.

Another possible explanation to why our methods segmented approach was less optimistic in these transfer combinations, may have been that we succeeded in identifying a segment in the development sample that had observations that were more similar to the observations after transfer, while the association between the predictors and the outcomes were also more similar between the segmented sample and the validation sample. If this was the case, the associations were still not similar enough to produce insignificant results in the segmented approach for these transfer combinations.

Another possible explanation to why our methods segmented approach was less optimistic in these transfer combinations, may have been that the number of misclassified validation observations were fewer. These misclassified validation observations are more similar to the development samples that were not included in the segmented samples. By having fewer of these observations, the segmented samples may have become better representatives of the validation samples in these transfer combinations.

## Methods predicting performance

There are several methods that propose to predict the performance of a prediction model more accurately in new patients that were not used to train the prediction model. These methods include different types of bootstrapping, cross-validation, and split-sampling methods (23-26). Even though these methods can predict performance more accurately in new patients, they are only capable of doing so in patients that are similar to those that were used to train the prediction model. These methods cannot provide information on how the prediction model will perform when the prediction model is transferred to another country (27).

If these methods were to be used to predict performance after transfer, one would expect to achieve performances that are similar to our naive approach performance. This is because our development sample accuracy being derived almost similarly to how cross-validation accuracies would be derived.

## Prediction model selection strategy

Although not being a usual way of selecting a prediction models, the way that we selected prediction models based on cross-validation should made the naive approach less optimistic. This is because we used fewer patients form 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 more difficult for our methods segmented approach to predict a performance after transfer that is as good as or better than the naive approach.

Even if we failed in selecting prediction models that were that were less optimistic in the naive approach, our results show that the naive approach were still better in most of our transfer combinations. This makes it less important if we failed in selecting prediction models that were less optimistic in the naive approach.

## 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 prediction model performance after transfer using unlabelled data. Such a method could simplify the process of implementing prediction models in health care and therefore indirectly improve 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 simulated samples were probably poor representatives of real samples as we simulated outcome using logistic regression models trained with the original samples. Logistic regression models that are trained with few events per variable, which some of our original samples had, has frequently been associated with poorer predictive performance (28). The simulated outcome 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, although no consensus exits for which method one should chose when selecting predictors for a prediction model, it has been recommended to include all available predictors to reduce the model from adapting too closely to the sample that is used for developing the prediction model (11). We followed this recommendation partially, as we excluded categorical predictors to simplify the analysis. This may have increased the adaptation to the sample used for developing the prediction model which may have reflected poorly on our method. Third, because we did not assess the number of patients from the segmented samples that were included in the training of the prediction models, we can only hypothesize how this may have reflected on our results. Fourth, because we did not assess the number of observations that were misclassified as development sample and validation sample by our propensity models, we can only hypothesize how this may have reflected on our results. Fifth, due to simplicity and the limitation of time, we chose to only assess the performance of our prediction models in terms of accuracy (number of correct predictions divided by total number of predictions) for one specific decision threshold. Our results are therefore limited to this decision threshold. Last, due to being limited by time, we decided to only test our method with one dataset with data from three different countries.

## Significance

Because our findings contradict our hypothesis, we cannot recommend that our methods segmented approach is used for predicting prediction model performance after transfer. At the same time, because the naive approach was shown to be mostly optimistic and the only option to our methods segmented approach, the study emphasizes the importance of validating prediction models with both predictors and outcome after transfer. This is until a new method is developed that can predict the performance after transfer more accurately than the naive approach.

### Equity

Equity was not an important part of this study as the aim was to only develop and test a method that predicts prediction model performance after transfer using unlabelled data. However, because this method could have simplified implementation of prediction models in health care, it is important to understand fairness in prediction models as they can be vulnerable to populations that have experienced human and structural biases (29). It is therefore important to incorporate fairness during prediction model design and development to ensure that all patients benefit from prediction models.

## Future studies

Because our study has several limitations that may have affected our method poorly, a future simulation study could try to improve on our study by reducing these limitations. This could be done by using sample sizes that are larger, by including all model predictors that are available from the dataset and by testing our method with several prediction models. Further improvements could also be done by assessing the performance of the prediction models in measures that is 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 that were included from the segmented samples may have affected our method, a future simulation 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 number of observations that were misclassified as development sample and validation sample by our propensity, a future simulation study could separate transfers based on this to gain a better understanding of the performance of our method and possible improvements to it.

# Conclusions

Although our method used unlabelled data to predict the performance 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 affected the method poorly, further studies are needed to gain a better understanding of our method’s true performance and possible areas of improvement.

# Contributions

The sample simulation was done by Martin Gerdin Wärnberg. The rest of the study including the statistical analysis and writing was done by me.

# 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. 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.

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

7. 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.

8. 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.

9. 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.

10. 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.

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

12. 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.

13. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. Journal of thoracic disease. 2019 Mar;11(Suppl 4):S574.

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.

29. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018;169(12):866–72.