BIOSTATISTICS - WRITTEN REPORT
2021/2022

Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients

Presentation group :
Céline HAJJAR
Dina EL ZEIN
Fabien MERCERON
David SOTO

Report written by :
David SOTO

Table of content

1 Introduction

Risk adjustment algorithms are important for the improvement of healthcare quality. Inter alia, these algorithms enable to predict hospitals patients' risk of mortality, which contributes to ameliorate medical care. A number of risk adjustment algorithms have been specifically developed to estimate and adjust for ICU patients' risk of mortality. The most documented mortality risk adjustment algorithms include the APACHE (Acute Physiology and Chronic Health Evaluation), the SAPS (Simplified Acute Physiology Score), and the MPM (Mortality Probability Model). Even though there are many validated and accepted risk adjustment algorithms available, utilization of these risk adjustment algorithms is low. There are two main explanations for the low use of these algorithms. First, although the algorithms APACHE-IV and MPM-III are in the public domain and available at no cost, to use these tools in practice typically requires the payment of licensing, implementation, and maintenance fees which many hospitals, if not all hospitals, cannot afford to pay. If the hospitals were to use this algorithms, they would have to dedicate a part of their financial budget for the use of these algorithms, which they cannot afford to do. Second, implementation of these algorithms often require critical care clinicians to engage in time-intensive collection and documentation of patient data that is not captured in typical critical care workflows nor is readily available in commonly used clinical information systems. Manual data collection is costly for any hospital and this is why no hospital even consider this data collection method. Inefficient clinical data collection methods have hampered the use of risk adjustment algorithms.
In response to federal incentive programs, hospital adoption of Electronic Health Record (EHR) systems has accelerated and today nearly 97% of hospitals have at least a basic EHR. With the widespread use of EHR systems, the hospitals have now the potential to obviate the need for risk adjustment algorithms that require manual data collection. On the other hand, the widespread use of EHR systems favors the utilization of risk adjustment algorithms that are based on automated data collection systems.
The article *"Development and Evaluation of an Automated Machine Learning Algorithm for In-Hospital Mortality Risk Adjustment Among Critical Care Patients"* presents and describes the development and evaluation of a risk adjustment algorithm using modern machine learning methods that rely on data that are routinely collected in the course of patient care and hospital operations. The risk adjustment model presented in the article uses data elements that can be extracted from clinical and administrative systems in an automated fashion and therefore requires no manual data collection or custom data entry. As we will see in the following parts, the risk adjustment algorithm presented in the study has demonstrated good performance in terms of discrimination, accuracy and calibration.

2 Materials

In this section, we detail the materials that were used to conduct the study.

2.1 Hospital and ICU selection

For the study, 79 care hospitals across the United States participated. The hospitals were selected based on four criteria. Firstly, the hospitals needed to have at least one ICU that meets the National Healthcare criteria, which is an adult critical care criteria. Secondly, the hospitals had not to be children's hospitals. Thirdly, the hospitals had to have a minimum of 100 ICU patients in the study period. Lastly, the hospitals had to use the Cerner Millenium EHR system. Out of the initial 79 hospitals that participated, only 53 hospitals met these criteria and were therefore retained to continue the study.

2.2 Patient selection

For the study, all inpatients 18 years or older discharged between the beginning of January 2014 and December 2016 that spent a portion of their hospital stay in an ICU were included in this analysis. Among these patients, unknown discharge disposition patients were excluded from the study. In total, 237,173 patients were retained for the study.

3 Methods

3.1 Feature selection

Clinical attributes available in the EHR and in administrative data amount to more than 15,000 total features. In particular, administrative data include thousands of features that could potentially be incorporated into a risk adjustment algorithm. Given the excessive amount of features, feature selection seemed a necessary step to pursue the study. Indeed, feature selection allowed to reduce the feature space to a more manageable size that would allow the authors to produce a risk adjustment algorithm that was feasible from both a computational and implementation standpoint. The feature selection process was carried out in three stages :

1. An environmental scan of existing risk adjustment algorithms
2. Consultation with a professional (in the study they consulted with a board-certified critical care physician).
3. Automated feature selection via Machine Learning models

Steps 1 and 2 of the feature selection process yielded a significantly reduced feature set that included only 215 clinical and administrative features. These features were carried forward to the last step, described in the following part.

3.2 Model development and evaluation

The objective of the model was to reliably estimate ICU patients' risk of in-hospital death. To this end, the authors used XGBoost, an open-source software library for machine learning, to build a model that could reliably estimate the risk of inpatient mortality. XGBoost models are comprised of thousands of relatively simple decision trees. These models are trained iteratively by combining

individual decision tree models to optimize on an evaluation metric, such as AUC. At each iteration, an additional decision tree is added to the aggregate of previously trained decision trees. However, each new decision tree takes into account errors made in the previous iterations. In this way the model learns its own shortcomings and introduces a new decision tree to address those shortcomings. XGBoost was also used for the final feature selection step. After the first 2 stages of feature selection, the 215 remaining features are entered into the supervised machine learning selection process. Relative influence of the model features was determined by evaluating the "gain", a value based on the number of times a feature is selected for splitting across the entire ensemble of decision trees. Gain is the increase in accuracy brought about by including a feature to the branches of the decision tree. The relative influence of a feature is averaged over all trees. Initially, all features were entered in to the model, several rounds of feature selection ensued. At each round, one third of the features (which correspond to the least influent features in terms of gain) was dropped from the model. This process was repeated until the authors observed substantial declines in model discrimination (in terms of AUC). This process reduced the number of features from 215 to 17. The 17 remaining features were retained in the final model.

Assignment to the training versus validation set was determined via simulation study in which the authors generated 200 random splits of the hospital population. For each random split, the authors calculated the ratios observed mortality/expected mortality for the training and validation sets. Finally, the split that minimized the difference in observed mortality/expected mortality between the training and validation sets was selected. The objective of this simulation exercise was to reduce the risk of bias that could occur if a disproportionate share of low or high mortality ICUs were assigned to the training or validation set. Figure 1 is a depiction of the process of determining the training and validation sets. Finally, the model was trained using patients from 36 of the 53 hospitals, and the patients from the 17 remaining hospitals composed the validation set. In overall, the training set was composed of 146,982 patients and the validation set of 90,191 patients.

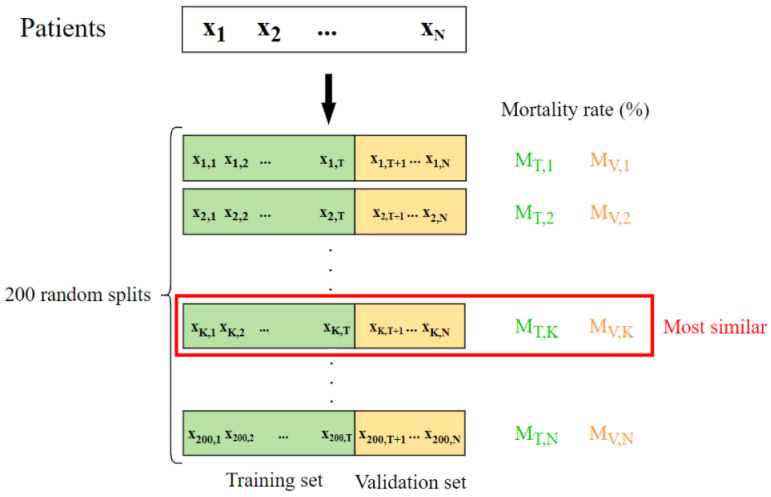

Figure 1: Assignment to the training versus validation set process

4 Results

Model performance in the study was assessed based on discrimination, accuracy, and calibration.

4.1 Calibration and Discrimination of the model

Calibration is about agreement between observed and predicted risk. Often, the Hosmer-Lemeshow test is used to assess the calibration of risk. However, this test can be unreliable when the sample size is large. For this reason, in the study, model accuracy and calibration were assessed using the adjusted Brier score and visual analysis of the calibration curves. The adjusted Brier score represents the percent reduction in deviation when using the RIPD model as opposed to assigning all patients a risk score equal to the average mortality for the entire population. A higher adjusted brier score indicates better model accuracy. The adjusted Brier score for the model is 52.8%, which is regarded as a good calibration by the authors because it compares favorably to the adjusted brier score obtained by the most commonly used risk adjustment algorithms. Moreover, visual analysis of the calibration curves, which are displayed in Figure 2, show good calibration of the model. The two curves in the figure are almost identical. Figure 2 shows close agreement between observed and predicted mortality rates across deciles of the risk of inpatient death (RIPD) score distribution. From the adjusted brier score and the calibration curves of the model, one can conclude that the model has good calibration.
In the study, model discrimination was evaluated based on the area under receiver operating characteristic curve (AUC). The AUC obtained for the model is 0.94, which is a good discrimination. Moreover, the model obtained a better AUC than the most commonly used risk adjustment algorithms such as APACHE (which has a AUC of 0.89) or MPM (which has a AUC of 0.81).

Thus, the mortality risk adjustment algorithm introduced by the authors has good calibration and good discrimination. These results compare favorably to the published performance statistics for the most commonly used mortality risk adjustment algorithms.

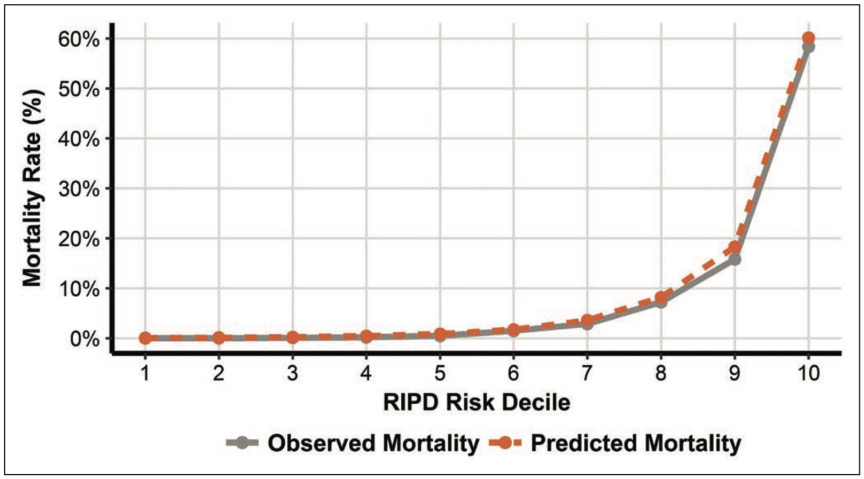

Figure 2: Plot of the calibration curves for visual analysis

5 Contribution of the study

The main contribution of the study is the development of a highly accurate risk adjustment algorithm with minimal cost, data- and licensing- wise. Moreover, the risk adjustment algorithm presented in the study addresses the problems encountered by the most commonly used mortality risk adjustment algorithms.
First of all, the risk adjustment model of the study addresses cost. Cost is by far the most significant barrier to adoption of risk adjustment algorithms in hospitals. This is because hospitals already spend a lot of money. For instance, the provision of healthcare in American ICUs costs already over $80 billion per year. Moreover, the introduction and diffusion of new technologies tends to push these costs higher. Hospitals, and especially ICUs, cannot afford to aggrandize their spendings. The risk adjustment algorithm presented in the article addresses this issue. The fact that this model is an automated algorithm based on freely available open-source software makes it available at no cost. This is a one of the chief improvements brought forth by this risk adjustment model.
Moreover, the risk adjustment algorithm presented in the study does not require any custom data entry. This model relies entirely on data elements that are likely to be accessible to hospitals that have electronic clinical and administrative systems, such as EHR systems. Given the widespread use of EHRs by the hospitals, the new model could be broadly and effectively adopted for risk adjustment purposes.

6 Discussion

The authors of the article have introduced a risk adjustment algorithm for ICU mortality. The study demonstrates the feasibility of developing an automated ICU risk adjustment algorithm using machine learning methods that only use highly available EHR and administrative data elements. Given the widespread adoption of EHR systems, the risk adjustment algorithm presented in the article could be widely and efficiently adopted. However, we notice some flaws in the development and evaluation of the model, which we describe in this section.

6.1 Absence of a test set

First of all, we notice the absence of a test set in the model development. This is probably the most striking flaw in the model's development. The authors do not use a proper test set in the development and evaluation of the machine learning model, they only use a train set and a validation set. The reasons for adopting this method is not explained in the article. One could imagine that such method was employed because of a limited amount of data available for the study, although 237,173 patients seems a fair amount of data to split into train-val-test sets. This approach is definitely not optimal for the development and the evaluation of the machine learning model. Indeed, it is always better to evaluate a model with new data that hasn't been seen by the model before. The main drawback of adopting such approach is that since validation set and test set have two different roles in the development and evaluation of a model, adopting this approach would have aftermaths on the model's performance. Neglecting one of these data sets would definitely lessen the performance for the model. The reason is that a validation set is used to provide an unbiased evaluation of a model fitted on the training set while still tuning the hyperparameters. On the other hand, the test set is used to provide an unbiased evaluation of the **final** model, fitted on the training set. The fact that the model was only evaluated on a validation set during its development implies that the evaluation of the model is not entirely reliable. Thus, by trying to merge the validation set and the test set, the

authors undermined a crucial step in the development and evaluation of the model, which results in an evaluation that is not totally reliable.

6.2 Risk of bias

Another flaw in the study regards the method used for the train-validation split. As explained earlier, the train-validation split was determined via simulation study in which the authors generated 200 random splits of the hospital population. The split that minimized the difference in observed mortality / expected mortality between the training and validation sets was retained. However, this method for splitting the data and the resulting train-validation split could be biased. When developing a machine learning model, it is always important to respect a certain homogeneity between the training data and the validation data. Yet, nothing in the study suggests that the homogeneity between the training data and the validation data was respected. In the study, there are 53 hospitals, each of which contains its own material and ICUs. The hospitals are located in 14 different states. The majority of these hospitals are located in metropolitan areas. However, there are some hospitals that are located in micropolitan areas, and even in rural areas. As we can imagine, there might be a noticeable difference between metropolitan hospitals and rural hospitals in terms of size, structure and materials. It is patent that the metropolitan hospitals are likely to possess more (and better) medical materials than the rural hospitals, more medical practitioners, and that the amount of patients and the diversity of diseases encountered in these hospitals are likely to be greater than the ones encountered in micropolitan hospitals. Not respecting the homogeneity between the training data and the validation data could distort the performance of the model. For instance, it could be the case that the train set is composed of only metropolitan hospitals and that the validation set contains all the micropolitan-rural hospitals. If that is the case, although the model was well trained on the training set, the model's performance is likely to be distorted in the evaluation of the model. The results obtained in the validation set would be biased. The reverse could happen as well. The train set could contain all the small hospitals and the validation set could have no rural hospitals. In this case the results would also be biased.
In order to ameliorate the study, the authors should find a way to ensure that there is the same proportion of metropolitan and micropolitan hospitals in the train set and in the validation set.

6.3 Ambiguous definition of the RIPD score

Another flaw in the article is the ambiguity in the definition of RIPD score. Throughout the article, the authors use the term RIPD score. However, the authors do not give a clear definition of this term. This does not help the reader to understand the paper. RIPD stands for Risk of Inpatient Death, and although RIPD score gives the impression of refering to the model's performance, the RIPD score seems to be the machine learning model's name, but the ambiguity persists because of a lack of an explicit definition. Aside from RIPD score, the authors also use the terms RIPD model and RIPD algorithm. Not providing a clear definition of these terms makes it harder for the reader to understand the paper.
This is definitely not the most insightful critic of the paper, yet it is relevant because providing the definitions of the terms used in the paper is one of the basic things that helps the reader to understand a study.