# Evaluating physician performance in critical care using healthcare service data. An ensemble approach of statistical and Machine Learning algorithms

**Eralda Gjika**[1], Amarildo Ceka[2], Belal Hossain[3]

[1] Carleton University, Ottawa   [2,3] University of British Columbia, British Columbia

## Background and Objectives

Recent developments and increase of information from various sources in clinical medicine have highlighted the need and relevance of prediction models. Facing with volume and variety of data in this domain surveys remain the main source of information. Using survey we manage to collect diverse information for patients and physicians who undertake important decisions at critical junctures. Recent studies have shown that machine learning models perform satisfactory when dealing with decision-making process of intensive critical care units (ICU) which main objective are:

- the **optimization** of human resources in order to,
- **minimize the mortality rates** that appear in the ICU wards and also,
- **minimize costs** for patients who are spending a considerable time in ICU.

Therefore, it is of high importance the evaluation of physician performance. Indicators which directly or indirectly affect physician performance may vary by: location, demography, human availability, financial capacities etc.

In this study, we investigate the information obtained from three different sources related to physician's performance. The main purpose is to combine this information in such an optimal way to obtain accurate evaluations of physicians' performance.

## Data

- **Physician had 360 feedback evaluation**

Linked to 7 domains (Medical Expert, Communicator, Leader, Advocate, Professional, Scholar, and Collaborator). Each question was scored on a 5-points likert scale. (Variable importance for target Q23)

- **Doctor characteristic data**

age, (2) background medical discipline before critical care training, (3) an administrative leader defined as a physician, (4) an education leader in the university, (5) university rank. (Variable importance for target M4)

- **Patient Participants (Patient Status and Outcome Scores)**

including demographic, clinical, and administrative information. (Variable importance for target ICU length of stay P9)

This study considers a dataset of 25 physicians which have 360 feedback evaluations completed anonymously by a random sample. A total of nine variables related to their characteristics are recorded from a survey on a 5-likert scale (756 observations). Another source of information is the patient status and outcome scores related to each physician (in total 2113 patients observed) are recorded and analyzed.

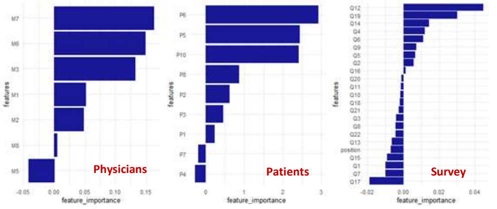


Fig.1 Feature importance (Random Forest)

## Pre-processing

- **Doctor characteristic data**

From the random forest model we observe the variable importance from physician characteristics. The graph shows the correlation between the domain where the physician received training (M7) and average overall score from survey in 2016 (M4). For M5 levels: *1 = 2 (8%) ; 2 = 14 (56%); 3 = 9 (36%)*

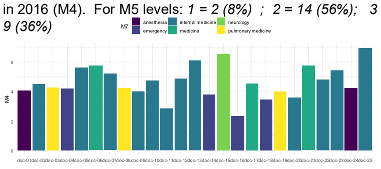


Fig.2 Physician training and overall score

- **Physician had 360 feedback evaluation**

From the random forest model we observe the variable importance are mainly related to questions on medical expert, professionalism and communication. Here we decided to create three new variables which measures the average performance of the physicians especially for these qualities and an overall average variable (average) which we compare with Q23 (overall score given ).

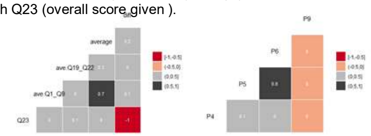


Fig.3 Correlation

- **Patient Participants (Patient Status and Outcome Scores)**

Using a random forest model with target ICU LOS (Length Of Stay) we observe the variable importance is high for SOFA and APACHE-II which are highly correlated (corr=0.8). Primary diagnosis (P10) and patient status at ICU discharge (P8) show a significant behavior related to physicians. It can be noticed that physician's faced with roughly more than three primary diagnosis of patients, encounter a lower proportion of death among their patients, and the vice versa.

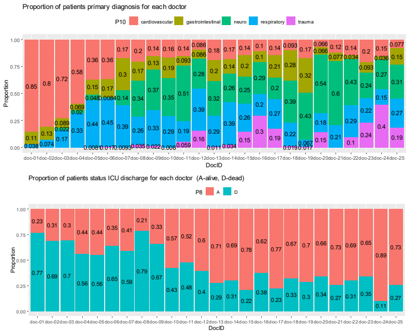


Fig.4 Proportion of diagnosis and status discharge of patients for each physicians

## Methods

We aimed at combining all three sources of information where doctor identification is the key used to connect the three datasets. We started our work by understanding each of the three dataset separately through descriptive statistics and graphs. Based also on previous work done related to the same issue we analysed the most important variables in each of the datasets and used them for performance classification.

- **Classification machine learning and cluster analysis**

Classification machine learning models such as random forest were used for variable importance. Cluster analysis and graphical representations were used to increase the accuracy in our model.

- For the **physician dataset** we used the average overall score from survey in 2016 (M4) and resident evaluation ranking group in 2016 (M5) as performance measures and the most important variable from our machine learning model were the rank of a physician (senior/junior) (M3) and the domain of the physician received training (M7).
- For the **patient dataset** we used as performance measurement the variables: **APACHE-II** score at admission (P5), **SOFA** score at admission (P6), **Status at ICU** discharge (P8) and **ICU length of stay** (P9). Referring to previous research done in the field and also from our analysis we found that for our data, the variables "SOFA" and "APACHE II" were significantly correlated (*correlation: 0.8). There is no significant correlation between SOFA and ICU LOS (Length Of Stay). (*We also used decision trees to investigate and justify our approach in using just SOFA instead of both in our model).
- On the **survey dataset** we added a quantitative variable as an average measure of the evaluations from Q1 to Q22. We noticed that there is no significant correlation between the overall score (Q23) and the scores given for each category separately (Scientific knowledge, Professionalism, Communication, Collaboration, Management). We also used average scores for each of the categories (obtained five new measurements) and investigated the impact of these scores on physician's performance.
- After performing, analyzing and modeling the performance of physicians based on each dataset in particular we further exploited these results by combining them and obtaining a final classification model.

## Results

- **Evaluation** In this study, we evaluated the ability of various machine learning algorithms to predict physician performance on ICU admission using data available from the SSD2022 competition. Based on our analysis of supervised and unsupervised methods, we classified from each source of information physicians based on a performance measure which was obtained from an investigation of each dataset. Then, we tried to use ensemble-based methods by combining the models from each source and finalizing the ranking. Qualitative judgment based on retrieval information from three sources was used to confirm the classification.

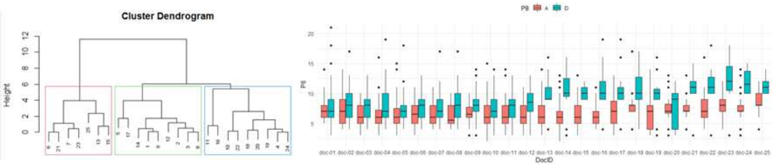
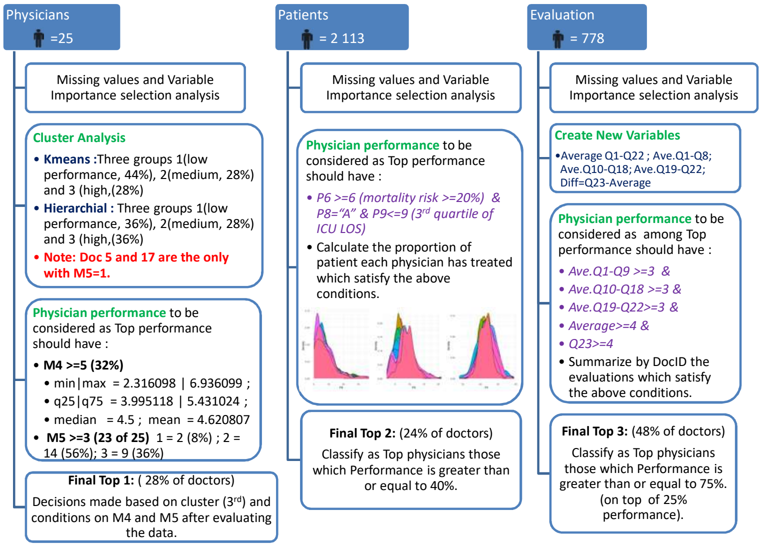


Fig.5 Physician classification using clusters

- **Schematic diagram representing the process of model development**



Physicians =25

Missing values and Variable Importance selection analysis

**Cluster Analysis**

- **Kmeans :**Three groups 1(low performance, 44%), 2(medium, 28%) and 3 (high,(28%)
- **Hierarchial :** Three groups 1(low performance, 36%), 2(medium, 28%) and 3 (high,(36%)
- **Note: Doc 5 and 17 are the only with M5=1.**

**Physician performance** to be considered as Top performance should have :

- **M4 >=5 (32%)**
  - min|max = 2.316098 | 6.936099 ;
  - q25|q75 = 3.995118 | 5.431024 ;
  - median = 4.5 ; mean = 4.620807
- **M5 >=3 (23 of 25)** 1 = 2 (8%) ; 2 = 14 (56%); 3 = 9 (36%)

**Final Top 1:** ( 28% of doctors)

Decisions made based on cluster (3rd) and conditions on M4 and M5 after evaluating the data.

Patients = 2 113

Missing values and Variable Importance selection analysis

**Physician performance** to be considered as Top performance should have :

- *P6 >=6 (mortality risk >=20%) & P8="A" & P9<=9 (3rd quartile of ICU LOS)*
- Calculate the proportion of patient each physician has treated which satisfy the above conditions.

**Final Top 2:** (24% of doctors)

Classify as Top physicians those which Performance is greater than or equal to 40%.

Evaluation = 778

Missing values and Variable Importance selection analysis

**Create New Variables**

- Average Q1-Q22 ; Ave.Q1-Q8; Ave.Q10-Q18; Ave.Q19-Q22; Diff=Q23-Average

**Physician performance** to be considered as among Top performance should have :

- *Ave.Q1-Q9 >=3 &*
- *Ave.Q10-Q18 >=3 &*
- *Ave.Q19-Q22>=3 &*
- *Average>=4 &*
- *Q23>=4*
- Summarize by DocID the evaluations which satisfy the above conditions.

**Final Top 3:** (48% of doctors)

Classify as Top physicians those which Performance is greater than or equal to 75%. (on top of 25% performance).

- **Ensemble model and final classification results** **(Top/Best physicians)**

| **Dataset used for Physicians Performance Evaluation** (High-Not High) | **Method** | | |
|---|---|---|---|
| | **Kmeans** | **Hierarchical Cluster** | **Performance based variables (M4 & M5)** |
| **Doctors** | **Gr 1:** 1, 2, 3, 4, 7, 8, 10, 12, 17, 22, 24 ( 44%) | **Gr 1:** 1, 2, 3, 4, 8, 10, 12, 22, 24 (36%) | **Top 1:**<br>6, 7, 13, 15, 21, 23, 25 (28%) |
| | **Gr. 2:** 9, 11, 14, 16, 18, 19, 20 (28%) | **Gr. 2:** 9, 11, 14, 16, 18, 19, 20 (28%) | |
| | **Gr.3:** 5, 6, 13, 15, 21, 23, 25 (28%) | **Gr.3:** 5, 6, 7, 13, 15, 17, 21, 23, 25 (36%) | |
| | **Top 1:** 6, 7, 13, 15, 21, 23, 25 ( 28%)<br>Note: Doc 5 and 17 are the only with M5=1. | | |
| **Patients** | **Top 2**: 15, 17, 21, 22, 24, 25 ( 24%) | | |
| **Evaluation/Survey** | **Top 3:** 3, 7, 8, 11, 15, 16, 17, 18, 19, 20, 22, 23 ( 48%) | | |
| **Final Top Physicians** | Final Top performance Doc ID: **15, 17, 21, 22, 23, 25** (28%) | | |

## Conclusions

- **Physician Performance Classification**
- **Overall** in our study we found that ensemble models have moderately better performance than using only statistical models.
- Machine learning random forest models we identified variables that had the highest impact on physician performance.
- We showed that **sex** and **education** of physicians, also **patient sex, age** and **primary diagnosis** were not important variables, and **average score** of each category played an important role in final classification.
- Using an ensemble machine learning model we obtained a final performance classification model of physicians. In our first approach the model classifies physicians in two categories: **"high"** and **"not-high"** and in the second approach we aim at classifying physicians in three categories: **"low", "medium"** and **"high".**
- **Challenges and Limitations**
  - We obtained final results for the first approach (top performance physicians). Further investigation on measures and their ranges should be considered. These ranges can be used in classification. And consideration of ordinal regression models will be our next approach .
  - The low number of physicians and their characteristics was one of the limitations and challenges we faced with. The different score levels were also challenging.
  - We build a **Shiny App** for a better performance of data manipulation and visualization. (scan the QR code for more)

## References and Acknowledgements

The authors acknowledge the support and supervision of Arsena Gjipali throughout the fruitful discussions and ideas on the study and further development of it.

References relevant to this work are available through the accompanying QR code.

Scan me!