2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

# Application of Machine Learning to Optimize Management of Children in Hospital with Lower Respiratory Tract Infection

Bastien Chapuis, Steve Cunningham, Donald Urquhart and Syed Ahmar Shah

*Abstract*— **Effective triage can help optimize the use of limited healthcare resources for managing paediatric patients with lower respiratory tract infection (LRTI), the primary cause of death worldwide for under 5 years old children. However, triage decisions do not consider medium to long term needs of hospitalized children. In this study, we aim to leverage data-driven methods using objective measures to predict the type of hospital stay (short or long). We used vital signs (heart rate, oxygen saturation, breathing rate, and temperature) recorded from 12,881 children admitted to paediatric intensive care units in China. We generated multiple features from each vital sign, and then used regularized logistic regression with 10-fold cross validation to test the generalizability of our models. We investigated the minimum number of recording days needed to provide a reliable estimate. We assessed model performance with Area Under the Curve (AUC) using Receiver Operating Characteristic. Our results show that each vital sign independently helps predict hospital stay and the AUC increases further when vital signs are combined. In addition, early prediction of the type of stay of a patient admitted for LRTI using vital signs is possible, even with using only one day of recordings. There is now a need to apply these predictive models to other populations to assess the generalizability of the proposed methods.**

## I. Introduction

Lower Respiratory Tracts Infections (LRTI) are the leading cause of childhood death worldwide, with 2.56 million deaths recorded including 808,920 aged under 5 years old in 2017 [1]. Current practice when assessing LRTI disease severity is principally based on a subjective clinical assessment, heavily dependent on training and experience. A widely used practice in clinical settings is to triage a patient during consultation i.e. allocate appropriate priority to a patient depending on the severity of their condition. It is critical to improve triage of patients to ensure that resources are appropriately allocated to reduce both over-diagnosis and under-diagnosis. Effective triage greatly affects the efficiency of health organisations by reacting to morbidity and reducing the number of deaths, especially during a crisis. Consequently, triage is important in clinical decision making to ensure that limited resources are appropriately utilised. However, triage focuses on the current state of the patient to define what are his/her needs and assigning treatment priority. Prediction of a patient's future needs in the medium to longer term is not considered in triage and management of resources is therefore limited to short-term. A tool able to anticipate patient's needs (e.g., time spent at the hospital) in the future based on early recorded data should be complementary to the triage process.

In order to triage patients with LRTI, vital signs (heart rate, breathing rate, oxygen saturation and temperature) are often recorded. These data could also be useful for medium to long-term estimations including the duration of the hospitalization of a patient. Nevertheless, the interpretation of vital signs for assessing severity is often subjective, leading to significant variance amongst trained professionals when assessing disease severity [2]. Some reasons for this difficulty are the challenges associated with accurately acquiring vital signs from children who may be agitated, the lack of standardised algorithms for combining multiple vital signs and the fact that children's physiology changes as they grow older (e.g. normal heart rate and breathing rate is different for children at different age groups [3]). Consequently, there is variance in healthcare resource use and patient outcomes [2].

In the context of paediatric patients admitted to hospital, predicting the duration of stay of children is potentially valuable as this can help ensure appropriate use of hospital resources including beds. In this study, we have used a large dataset of 12,881 paediatric patients admitted to hospital to investigate whether we can use routinely collected vital signs (heart rate (HR), breathing rate (BR), oxygen saturation ($SpO_2$) and temperature) to help predict the duration of hospital stay with a minimal amount of data (including triage data and patient's monitoring data during hospital stay).

## II. Method

Figure 1 provides an overview of the methods in our study. We will first describe the Paediatric Intensive Care (PIC) dataset, and then explain the methods used to prepare (data preprocessing, normalization, feature extraction, and labelling and splitting) and analyze this dataset.

### A. Study Dataset

The PIC data, available on PhysioNet to bona fide researchers [4], were collected between 2010 and 2018 at The Children's Hospital, Zhejiang University School of Medicine in 5 different intensive care units (ICU). These five different ICUs (General, Paediatric, Surgery, Cardiac and Neonatal) had a combined capacity of 119 beds [4]. The dataset contains laboratory information, observations from nurses and physicians including diagnosis, hospital electronic medical records, and demographic information for each patient. To ensure that our results are broadly applicable in several countries, we only extracted routinely collected data from the database for subsequent modelling. This included patient's vital signs, age (extracted

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

using patient's birth date), physician's diagnosis and duration of hospital stay.

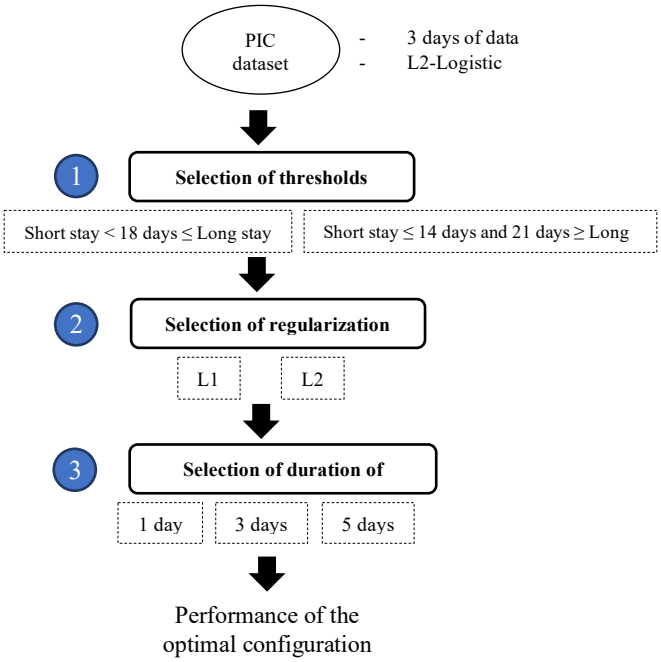


*Figure 1: Overview of methods employed*

### B. Data Preprocessing

The first step was to create a list of diseases pertaining to LRTI. In the study dataset, patient's pathologies are described with ICD codes with corresponding description text. With expert consensus, we defined relevant search terms to identify a subset of patients with LRTI. The search terms used were: "bronchiolitis", "LRTI", "pneumonia", "bronchitis", "empyema" or "Acute upper respiratory infection".

Owing to the large amount of missing data, we subsequently derived four datasets, one each for one of the four vital sign: HR, BR, $SpO_2$ and temperature. For each vital sign, we defined limits to ensure that any outliers are removed. The outlier limits were defined to remove physiologically impossible values. These limit were: ($HR \notin [4\,;\,300]$, $RR \notin [4\,;\,100]$ , $SpO_2 \notin [40\,;\,100]$ and $Temperature \notin [25\,;\,45]$). We further refined the outlier removal for HR and BR by calculating mean (μ) and standard deviation ($\sigma$) and removing any values that were outside the range of $[\mu - 3\sigma\,;\,\mu + 3\sigma\,]$.

Finally, only patients with enough records per vital signs were selected. Three different duration of collections were analyzed: 1 day, 3 days and 5 days requiring respectively at least 3, 10 and 20 records.

### C. Data Normalization

After cleaning, HR and BR were normalized to ensure that the values are comparable across database. This is because the normal values of HR and BR vary from birth to adult age [3]. These variables were expressed as a distance from the reference value that a same aged healthy child would have. The distance from the normal is in $\sigma$ and ensures that a valid comparison is feasible (e.g. between a 15-year-old patient and a 3-day-old baby [3]).

$SpO_2$ was also modified fixing every value above the mean ($\mu_{SpO2}$ =97.25%) as unlike other vital signs, an $SpO_2$ above $\mu_{SpO2}$ is not abnormal.

### D. Feature extraction

There was considerable variability in how often vital signs were recorded. This meant that the vital signs were irregularly sampled and there was no specific pattern of recording, e.g., some patients had more (and/or more frequently measured) vital sign measurements than others. We, consequently, devised a feature extraction method that can deal with irregular sampling and variable number of records.

For each vital sign, we extracted five different features: the μ (average of recorded values), the trend (the way values change over time), the size (total number of records), the variability of vital sign values ($\sigma$ of measurements values) and the variability in the recording frequency ($\sigma$ of time duration between each recording time). For illustration, Figure 2 shows a patient record where we have extracted these five features.

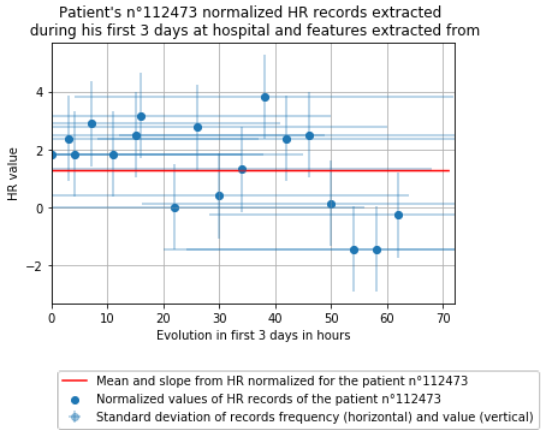


*Figure 2: Features extracted from cleaned and normalized data: the mean, slope, variation of value, variation of time and number of records.*

These features were selected to lose minimal amount of information and give algorithms the possibility of distinguishing patients with stable state from patients getting worse or better, and patients with few records (e.g., 10 records over 3 days) from patients with several recordings (e.g., >100 records over 3 days).

The mean of vital signs recorded corresponds to the sum of every record divided by the number of records. It does not take in consideration the time gap between 2 records, the total number of records or the gap between the lowest and the highest value. The slope is independent of each individual vital sign value. It represents their trend (increasing, decreasing or no change). The same trend would be noted for a patient whose $SpO_2$ increases from 70% to 71% as a patient whose $SpO_2$ increases from 98% to 99%. In our study, we estimated the trend of vital sign records by fitting a straight line and extracting the slope of the line as a measure of the trend. Finally, the standard deviation, $\sigma$, provides an estimate of the variation of values around the mean.

### E. Labelling and splitting of data

We defined two types of stay: "Short stay" and "Long stay" based on the number of days spent at the hospital and tested two thresholds to define short and long stays. Firstly, we used the median number of days spent at the hospital (18 days) allowing equal splitting between the two types of stay. Subsequently, with

expert consensus, we defined short stay as up to 2 weeks long and long stay as over 3 weeks ("Short stay" ≤14 days and "Long stay" ≥21 days).

### F. Logistic Regression

In this study, we have used logistic regression, a machine learning algorithm that is typically used as a benchmark. It is a very popular algorithm because of its simplicity [5] and interpretability [6].

In the context of the study, we have 2 types of stays ("Short" and "Long") that we will classify (the output) using a fixed number of input parameters, dependent on the dataset. Mathematically, we consider the input as a vector $X=\{x_0, x_1, x_2, \dots, x_n\}$ with $x_0 = 1$ and $x_1, \dots, x_n$ the n parameters (5 per vital signs).

The 2 possible outputs are converted into a number, generally using 0 and 1, that the model will aim to approach using a combination of input parameters $X$ and weights $P$ (1, 2). P is the vector of weights $\{p_0, p_1, p_2, \dots, p_n\}$, attributed to each input in $X$. The prediction of the model is defined using the sigmoid function $g$ (3) on the scalar product of $P$ and $X$ (1).

$$P \cdot X = p_0 * x_0 + p_1 * x_1 + p_2 * x_2 + \cdots + p_n * x_n \quad (1)$$

$$h_p(X) = g(P \cdot X) \quad (2)$$

$$g(y) = \frac{1}{1 + e^{-y}} \quad (3)$$

During the training phase, the parameters of the model are tuned such that the total "cost" of making errors is minimized. This cost is calculated with the cost function (4).

$$C(P) = -\frac{1}{m}\left(\sum_{i=1}^{m} y^{(i)} * log\left(h_p\left(X^{(i)}\right)\right) + \left(1 - y^{(i)}\right) * log\left(h_p\left(X^{(i)}\right)\right)\right) \quad (4)$$

with $m$ the number of cases in the training set.
This function returns high values when the predicted value $h_p\left(X^{(i)}\right)$ and the expected value $y(i)$ are different and low values when they are similar.

### G. Cross validation

A 10-fold cross validation strategy was adopted during this project to ensure generalization [7]. It consists of splitting the data into 10 parts (called folds), and then iterating 10 times over all folds, each time using a different fold for testing while using the remaining folds for training. Finally, the overall performance of the algorithm was assessed by taking the first and second order statistics of the performance of each of the ten models.

### H. Regularization

All logistic regression model trained during this project were optimized with regularization. This concept aims to limit overfitting penalizing complex solution in favor of the simple ones [8][9].

Two regularization methods are mainly used: L1-regularization, also called Lasso (Least Absolute Shrinkage and Selection Operator) regression and L2-regularization also called Ridge regression [8]. L1 and L2 regularizations are implemented as part of the logistic regression by adding, respectively,

$$R_{Lasso}(P) = \lambda \sum_{j=0}^{n} |p_j| \quad (5)$$

or

$$R_{Ridge}(P) = \lambda \sum_{j=0}^{n} p_j^2 \quad (6)$$

to the cost function $C(P)$ with $\lambda$ a parameter to modulate the impact of the regularization on the cost function and $n$ the number of weight and size of $P$ [10]. We can see from Lasso and Ridge formulas (5, 6) that the main difference between both is the power of the weight terms.

### I. One Standard Error rule

The Lasso regression is also used to reduce complexity of a model reducing weakest $P$-weight coefficients to 0. However, the efficiency of this method directly depends on the $\lambda$ value. The 1 Standard Error (1SE) rule is a heuristic method to select the optimal $\lambda$ for Lasso regularization to get best performances with least number of parameters [11]. Applying the 1SE rule consists of testing different values of $\lambda$ and calculating, for each of them, the resulting average Mean Squared Error (MSE). Then, the lowest MSE amongst all calculated is selected, the standard error is estimated, and the highest $\lambda$ value with MSE below the $Lowest\ MSE + Standard\ Error$ threshold is considered as the optimal $\lambda$ value. Coefficients of the model are weights attributed to each input parameters. A weight equal to 0 means the corresponding parameter is not used for the prediction.

### J. Classification and metrics

We used two different thresholds to define short stay and long stay. For both the thresholds, we used L2-regularized logistic regression and compared the performance of the algorithms. We subsequently compared the influence of L1 and L2 regularizations on the performance of logistic regression-based classifier. We also sought to determine the minimum duration of data needed to predict whether a hospital stay will be long or short. For this step, we trained logistic regression models using 1 day, 3 days and 5 days of data and compared the performance of the algorithms. Finally, Lasso regression was also applied to rank features according to their importance and to identify the most useful parameters that can help predict the type of hospital stay.

We used the area under the Receiver Operator Characteristic (ROC) curve to compare the performance of models [11]. The ROC is a widely used technique that can help assess the performance of a classifier and takes both the true positive rate and the false positive rate of a model into consideration.

## III. Results

The PIC dataset was collected from 12,881 children (7,366 boys and 5,515 girls) each of whom presented one or multiple times to one of the ICU departments between 2010 and 2018. Out of those, 971 died during their stay. Details of vital sign values of patients are provided in Table 1.

Our search strategy, by expert consensus, identified 1,194 patients with LRTI from the database, who were then selected for subsequent analysis.

Performances of the logistic regression algorithm with L2 regularization and the initial threshold (18 days) to separate "Short stays" from "Long stays" showed that each vital sign independently helped to predict duration of hospitalizations with an AUC between 0.642 for temperature, 0.653 for $SpO_2$, 0.701

for BR, and 0.727 for HR. When all the vital signs were combined, the AUC increased to 0.782. Table 2 provides the AUC with the associated confidence intervals for all possible combination of vital signs when separating long stay (≥18 days) from short stay (<18 days).

We were also able to differentiate between a long stay and a short stay where "long stay" was now defined as ≥21 days, and "short stay" was defined as ≤14 days. In this case, we found the AUC for using a single vital sign varied from 0.667 to 0.736. However, a model that used HR, BR and $SpO_2$ resulted in the best performance (AUC=0.814).

*Table 1: Statistics of vital signs in the dataset*

| Vital sign | Heart rate | Respiratory rate | Oxygen Saturation | Temperature |
|---|---|---|---|---|
| **Number of records** | 462,770 | 709,606 | 287,310 | 445,475 |
| **Number of admissions** | 10,856 | 10,823 | 8,078 | 10,521 |
| **Number of patients** | 10,401 | 10,351 | 7,803 | 10,049 |
| **Median (5%-95%)** | 131 (80 - 167) | 34 (20 - 54) | 97 (83 - 99) | 36 (36 - 38) |
| **$\mu^a$ ($\sigma^b$)** | 128 (27) | 36 (12) | 96 (6) | 37 (1) |

[a] Mean

[b] Standard deviation

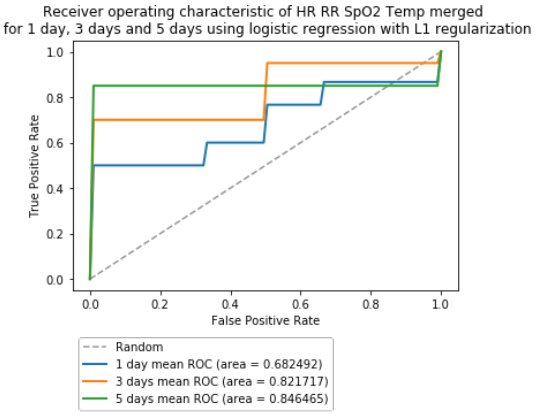


*Figure 3: ROC of L1-logistic regression for 1 day, 3 days and 5 days of data merging all vital signs.*

We did not find a significant difference between the influence of L1 and L2 regularizations on the performance of logistic regression-based classifiers. The performances of L1 and L2 regularization were similar for most of models except for the merge of HR, BR and $SpO_2$ ($AUC_{L1}$=0.814 and $AUC_{L2}$=0.715). In addition, we also looked at how many days of data would be required to correctly identify the duration of stay. We compared 3 models using L1 regularization and all vital signs, attempting to differentiate long stay (≥21 days) from short stay (≤14 days). Figure 3 provide the ROC curves comparing using data from day 1 only, for first 3 days and first 5 days. The results suggest that we can differentiate between long stay and short stay using vital signs from the first day of admission, with an AUC of 0.682. However, the accuracy of the prediction increases when data from more days are used (0.822 for 3 days, and 0.846 for 5 days). Lastly, when we ranked the importance of features with LASSO, we found that the most useful features were the variability of frequency of recording, and the number of records.

## IV. Conclusion and Future Work

Our study shows that early prediction of type of stay (short vs long) of a patient admitted for LRTI using vital signs recordings is possible and does not require collection of large amounts of data, nor collection over long periods of time. At least 3 records per vital signs collected during the first 24 hours after admission can predict length of stay with better than chance level accuracy. We also found that healthcare delivery-associated features were the most useful for predicting the length of stay. These results suggest that re-assessing patients at 24 hours after admission may be useful for risk stratifying patients to assign appropriate treatment priority. There is now a need to apply these predictive models to other populations to assess the generalizability of the proposed methods.

*Table 1: ROC using L2-logistic regression on 3 days of data split in 2 classes: Short stay (<18 days) and long stay (≥18 days), and the number of patients in each class*

| Model | AUC mean (95% CI) | Short stays | Long stays | Total |
|---|---|---|---|---|
| **HR** | 0.727 (0.583 - 1) | 60 | 39 | 99 |
| **BR** | 0.701 (0.596 - 0.812) | 202 | 214 | 416 |
| **$SpO_2$** | 0.653 (0.517 - 0.903) | 151 | 171 | 322 |
| **Temperature** | 0.642 (0.518 - 0.841) | 144 | 160 | 304 |
| **HR & BR** | 0.750 (0.700 - 0.950) | 57 | 39 | 96 |
| **HR & $SpO_2$** | 0.859 (0.667 - 1) | 30 | 27 | 57 |
| **HR & Temperature** | 0.719 (0.550 - 1) | 53 | 37 | 90 |
| **BR & $SpO_2$** | 0.738 (0.615 - 0.926) | 141 | 168 | 309 |
| **BR & Temperature** | 0.746 (0.656 - 0.927) | 133 | 153 | 286 |
| **$SpO_2$ & Temperature** | 0.688 (0.558 - 0.880) | 89 | 130 | 219 |
| **HR & BR & $SpO_2$** | 0.726 (0.444 - 1) | 28 | 27 | 55 |
| **HR & BR & Temperature** | 0.749 (0.667 - 1) | 52 | 37 | 89 |
| **HR & $SpO_2$ & Temperature** | 0.765 (0.500 - 1) | 24 | 26 | 50 |
| **BR & $SpO_2$ & Temperature** | 0.726 (0.583 - 0.865) | 88 | 129 | 217 |
| **HR & BR & $SpO_2$ & Temperature** | 0.782 (0.667 - 1) | 24 | 26 | 50 |

## References


[1] C. E. Troeger *et al.*, "Quantifying risks and interventions that have affected the burden of lower respiratory infections among children younger than 5 years: an analysis for the Global Burden of Disease Study 2017," *Lancet Infect. Dis.*, vol. 20, no. 1, pp. 60–79, 2020, doi: 10.1016/S1473-3099(19)30410-4.

[2] T. A. Florin *et al.*, "Reliability of examination findings in suspected community-acquired pneumonia," *Pediatrics*, vol. 140, no. 3, Sep. 2017, doi: 10.1542/peds.2017-0310.

[3] S. Fleming *et al.*, "Normal ranges of heart rate and respiratory rate in children from birth to 18 years of age: A systematic review of observational studies," *Lancet*, vol. 377, no. 9770, pp. 1011–1018, 2011, doi: 10.1016/S0140-6736(10)62226-X.

[4] X. Zeng *et al.*, "PIC, a paediatric-specific intensive care database," *Sci. Data*, vol. 7, no. 1, pp. 1–8, Dec. 2020, doi: 10.1038/s41597-020-0355-4.

[5] S. C. Bagley, H. White, and B. A. Golomb, "Logistic regression in the medical literature: Standards for use and reporting, with particular attention to one medical domain," *J. Clin. Epidemiol.*, vol. 54, no. 10, pp. 979–985, 2001, doi: 10.1016/S0895-4356(01)00372-9.

[6] S. Dreiseitl and L. Ohno-Machado, "Logistic regression and artificial neural network classification models: A methodology review," *J. Biomed. Inform.*, vol. 35, no. 5–6, pp. 352–359, Oct. 2002, doi: 10.1016/S1532-0464(03)00034-0.

[7] S. Barratt and R. Sharma, "Optimizing for Generalization in Machine Learning with Cross-Validation Gradients."

[8] J. Goeman, R. Meijer, and N. Chaturvedi, "L1 and L2 Penalized Regression Models," 2018.

[9] B. Ghojogh, B. Ca, M. Crowley, and M. Ca, "The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial."

[10] J. Daniel and J. H. Martin, "Speech and Language Processing," 2020.

[11] T. Hastie, R. Tibshirani, and J. Friedman, "Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction."