Dear Editor and Reviewers:

We highly appreciate the referees’ detailed and valuable comments on our manuscript, titled “Thirty-day hospital readmission prediction model based on common data model with weather and air quality data” The suggestions were quite helpful for us and we incorporated them into the revised paper.

On behalf of my co-authors, I would like to clarify some of the points raised by the reviewers. We hope that you will be satisfied with our responses to your comments and with the revisions to the original manuscript.

Thank you and best regards!

Yours Sincerely,

Jinwook Choi and Sooyoung Yoo

2021-10-08
Reviewer's Comment to the Author

Title: Thirty-day hospital readmission prediction model based on common data model with weather and air quality data
Corresponding Author: Jinwook Choi and Sooyoung Yoo
Authors: Borim Ryu, Seok Kim, Sooyoung Yoo, Jinwook Choi

### Further Editorial / Peer-reviewer Comments:

Reviewers’ comments:

Reviewer 1: In this study, four machine learning models were developed to predict hospital readmissions based on the data of healthcare, meteorological conditions, and air pollution. This manuscript is well written in general. Please address the following comments.

1. Abstract. (1) Please add quantitative results to support the main conclusion of this study. (2) While AdaBoost exhibited the best performance for musculoskeletal diseases, GBM showed the best performance for other diseases. Why is only AdaBoost mentioned in the abstract?

Response: Thank you for the comment. Following your advice, we revised the overall content of Abstract: (1) Add quantitative results, (2) Corrected the results of musculoskeletal disorders.

2. This study only included the adults over 65 years old. Please explain why. Is it possible to include people under 65 years old?

Response: Thank you for the comment. In our previous study, we compared the readmission prediction model performance in a group of patients aged 65 years or older and a group of children aged 19 years and younger. Because the performance of the readmission prediction model for pediatric patients under 19 years of age was relatively lower than that of the elderly patients aged 65 years or older, this study was conducted on the patient group 65 years and older, which had relatively good predictive performance.

3. Figure 1. (1) It is a bit surprising to see that the external validation results are better than the internal validation results. Please explain why. (2) This figure only shows the results for AdaBoost and Decision tree. Please add the results for RF and GBM.

Response: Thank you for the comment.

(1) I agree with your opinion and the external validation results of our research results are
surprising. The original purpose of external validation is to verify how generalized and interpretable this model can be for performance evaluation of the developed model. In our results, it is estimated that this result occurred because the size of the data for which the model was developed and internal validation was performed and the size of the source data subjected to external validation were very different. Since the results were better in the verification process of the model for larger data, I think that the model created in this study is well made enough to generalize.

(2) RF and GBM result plots are additionally attached in the Figure 1. We revised the figure caption as follows: ROC curves for internal and external validation sets for tree-based machine learning models derived by clinical covariates and w-score.

4. Hyperparameter optimization is critical for the machine learning models. Based on my own experiences, GBM usually outperformed the other three models if hyperparameter optimization was properly performed. Please show the details of the hyperparameter optimization.

**Response:** Thank you for the comment. In our study, of course, we performed hyperparameter optimization. Details are additionally described in the main text. And the parameter values for each model in our study are as shown in the table below.

Summary of parameter values in each model

<table>
<thead>
<tr>
<th>Models</th>
<th>Parameters</th>
<th>Values</th>
<th>Parameter Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT</td>
<td>classWeight</td>
<td>“Balance” or “None”</td>
<td></td>
</tr>
<tr>
<td></td>
<td>maxDepth</td>
<td>10</td>
<td>The maximum depth of the tree</td>
</tr>
<tr>
<td></td>
<td>minImpuritySplit</td>
<td>10^-7</td>
<td>Threshold for early stopping in tree growth. A node will split if its impurity is above the threshold, otherwise it is a leaf</td>
</tr>
<tr>
<td></td>
<td>minSamplesLeaf</td>
<td>10</td>
<td>The minimum number of samples per leaf</td>
</tr>
<tr>
<td></td>
<td>minSamplesSplit</td>
<td>2</td>
<td>The minimum samples per split</td>
</tr>
<tr>
<td>RF</td>
<td>max depth</td>
<td>4, 10, 17</td>
<td>Max levels in a tree</td>
</tr>
<tr>
<td></td>
<td>mtries</td>
<td>-1 = square root of total</td>
<td>Number of features in each tree</td>
</tr>
<tr>
<td></td>
<td>features,5,20</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>---------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td><code>ntrees</code></td>
<td>500</td>
<td>Number of trees</td>
<td></td>
</tr>
<tr>
<td>ADA learning rate</td>
<td>1</td>
<td>Learning rate shrinks the contribution of each classifier by learning_rate. There is a trade-off between learningRate and nEstimators</td>
<td></td>
</tr>
<tr>
<td>n estimators</td>
<td>4</td>
<td>The maximum number of estimators at which boosting is terminated</td>
<td></td>
</tr>
<tr>
<td>GBM learning rate</td>
<td>0.005,0.01,0.1</td>
<td>The boosting learn rate</td>
<td></td>
</tr>
<tr>
<td>earlyStopRound</td>
<td>25</td>
<td>Stopping after rounds without improvement</td>
<td></td>
</tr>
<tr>
<td>max depth</td>
<td>4, 6, 17</td>
<td>Max levels in a tree</td>
<td></td>
</tr>
<tr>
<td>minRows</td>
<td>2</td>
<td>Min data points in a node</td>
<td></td>
</tr>
<tr>
<td>ntrees</td>
<td>100,1000</td>
<td>Number of trees</td>
<td></td>
</tr>
</tbody>
</table>

5. Please add model interpretation results such as variable importance, partial dependence plots, or SHAP to explore the impact of environmental factors and healthcare-related predictors.

**Response**: Thank you for the comment. Although we have not yet checked the feature importance results using something like SHAP, we checked the results for the effect of individual covariates in each model with the results presented in the supplementary material. Please refer to our supplementary material.

6. “KMA and Air Korea data needed to be preprocessed into postal zip codes owing to the varying levels of location information granularity.” Please clarify the method of preprocessing.

**Response**: Thank you for the comment.

LOCATION_ID in CDM DB has an address identifier based on the postal code address system. For example, LOCATION_ID for Jongno-gu, Seoul does not match SNUH LOCATION_ID and SNUBH LOCATION_ID value. Therefore, it is necessary to check the details of the LOCATION table of each institution CDM DB first. Meteorological data of KMA and air
environment data of Air Korea are recorded at each measuring station across the country, respectively. KMA data is divided into cities/metropolitan cities/provinces, and Air Korea data is based on a smaller unit, street address. Therefore, we first integrated KMA data and Air Korea data with the same granularity, and performed preprocessing by finding the postal code for the integrated address and matching it with the patient's residence address.

7. “based on the weather and air quality feature set using the patient-level prediction R package developed by OHDSI”. Please add citation(s).

**Response**: Thank you for the comment. We added the citation in the Reference.


**Reviewer 2**: Ryu and colleagues develop machine learning models to predict unplanned hospital readmissions within 30 days of discharge. Machine learning is a promising tool and its implementation in the clinical field is worth further investigation. However, in this manuscript, there are several issues that should be addressed before getting published in *Scientific Reports*.

**Specific Comments**

1. Introduction: Please describe and discuss previous modeling studies related to your topic and demonstrate why this study will contribute to this field.

**Response**: Thank you for the comment. In this study, the W-score of an individual patient was obtained by adding up the forecast values for each weather element for 7 days from the date of discharge so that the weather forecast data of the Korea Meteorological Administration could be used at the time of discharge. Since this is to predict the re-hospitalization of this patient at the time of discharge, it was designed in consideration of receiving the weather forecast data for the next 7 days for the patient's address and using it for clinical decision-making.
2. Methods: The choice of hyperparameters in machine learning models can greatly impact model performance. The hyperparameter tuning process needs to be clearly described in the Methods.

**Response:** Thank you for the comment. In our study, of course, we performed hyperparameter optimization. All possible combinations of the hyper-parameters are included in a grid search using cross-validation on the training set. Ten-fold cross-validation used to select the optimal hyper-parameter and internal validation. The hyper-parameters that lead to the best cross-validation performance will then be chosen for the final model.

3. Results: The hyperparameters used in the final models should be presented.

**Response:** Thank you for the comment. The parameter values for each model in our study are as shown in the table 5.

4. Results: The validation results for all groups of diseases & all machine learning models should be presented. The validation results for all groups of diseases & all machine learning models is presented as supplementary materials.

**Response:** Experimental result file was also uploaded according to our manuscript as supplementary materials already.

5. There are some language mistakes in the manuscript. Please check and revise.

**Response:** Re-edited for English proofreading.

**Reviewer 3:** The authors developed a model to predict unplanned hospital readmissions within 30 days of discharge based on the common data model considering weather and air quality factors. External validation demonstrated that the model based on weather and air quality factors is transportable, which might serve as reference values for further study. However, there are some questions/suggestions for the manuscript should be noted.
Major comments

1. Abstract section. Please present more data results instead of unnecessary words.

**Response:** Thank you for the comment. We revised the Abstract as follows: “Although several studies have attempted to develop a model for predicting 30-day re-hospitalization, few attempts have been made for sufficient verification and multi-center expansion for clinical use. In this study, we developed a model that predicts unplanned hospital readmission within 30 days of discharge; the model is based on a common data model and considers weather and air quality factors, and can be easily extended to multiple hospitals. We developed and compared four tree-based machine learning methods: decision tree, random forest, AdaBoost, and gradient boosting machine (GBM). AdaBoost-based demonstrated the highest AUC performance of 84.14 ± 0.015 for musculoskeletal disease group, and GBM showed the best performance for other diseases. Further, PM10, rainfall, and maximum temperature were the weather and air quality variables that most impacted the model. In addition, external validation has confirmed that the model based on weather and air quality factors has transportability to adapt to other hospital systems.”

2. Method section. Why the 30 days of discharge were used?

**Response:** Thank you for the comment. Broadly defined, a hospital readmission is when a patient who had been discharged from a hospital is admitted again to that hospital or another hospital within a specified time frame. The original hospital stay is often called the "index admission" and the subsequent hospital stay is called the "readmission." Different time frames have been used for research purposes, the most common being 30-day, 90-day, and 1-year readmissions. In the United States, as part of the Affordable Care Act at the national level, institutional mechanisms have been established, such as providing incentives for each hospital based on the indicator of readmission within 30 days of discharge. Since then, various studies have been conducted on unplanned hospital readmissions within 30 days, and many countries are using it as an indicator of quality of care. From this point of view, our study also set the goal of predicting readmission within 30 days to make a useful system applied to actual clinical practice.
3. Method section. Five pollutants including PM10, PM2.5, SO2, NO2, and O3 from all general monitoring stations were collected. However, the models just contain PM10 instead of other air pollutants. Please clarify it.

Response: Thank you for the comment. In this study, while calculating the W-score, the Korea Meteorological Administration's criteria for issuing meteorological advisories were referred to, and in this process, only PM10 was used among the atmospheric environment data. In Korea, fine dust forecasts and ultra-fine dust forecasts are divided into four levels of good, normal, bad, and very bad based on PM10 and PM2.5, respectively. PM10 was reflected in the W-score because there were many missing values of PM2.5 in the source data, and the multicollinearity relationship between PM10 and PM2.5 was confirmed.

Minor comments

1. The article lacks line number and this is very strange.

Response: Thank you for the comment. The manuscript is revised.

2. Result section. All the tables should be presented in three-line tables of scientific papers

Response: The entire manuscript was revised again in accordance with the journal formatting regulations.

3. Language needs to be polished for the manuscript. Avoid too many long and complex sentences in the manuscript

Response: Thank you for the comment. English proofreading again.