Weight-based Integrated Predictive Modelling of Horizontally Distributed Data Under Privacy-Preserving

Ji Ae. Park, MS[1], Hae Reong. Kim, MS[1], Kyu Yong. Lee[2], MinDong. Sung, MD [1], Jeong Hoon. Lee, PhD [1, 3], Dae-Il. Yang [1], Se Hee. Oh [1], Yu Rang.Park, PhD [1]

[1]Department of Biomedical Systems Informatics, Yonsei University College of medicine, Seoul, South Korea [2]Yonsei University College of medicine, Seoul, South Korea [3]Lunit Inc., 175 Yeoksamro, Gangnam-gu, Seoul 06247, Seoul, South Korea

Background

Health care data exists as distributed forms in numerous organizations. It is important to share the distributed data to obtain more various characteristics or information of patients. If the distributed data can be used to obtain various patient information, it would provide a better insight for a research[1]. Despite the benefits of utilizing distributed data, it is difficult to share the health care data of individuals, with confidential characteristics, because of ethical, administrative, legal, and political barriers[2].

In order to integrate horizontally distributed data, it is needed to standardize the data through Common Data Model(CDM), but way to analyze the distributed data is another important issue. The goal of this study is to develop an prediction model based on multi-institutional data without sharing patient-level data.

Methods

Suppose that there are three institutions (A, B, and C) and sample size of each institution is equal to n. The process of weight-based integrated model is as follows.

The data of each institution are randomly split into two part, $\mathrm{Z}^{(1)}$ and $\mathrm{Z}^{(2)}$. The first part, $\mathrm{Z}^{(1)}$, is to estimate a predictive model and the second part, $\mathrm{Z}^{(2)}$, is to measure the performance of the predictive model. Generate m pairs of $\mathrm{Z}^{(1)}$ and $\mathrm{Z}^{(2)}$, respectively, in each institution. Let i , i=1,…,m, denote the number of a pair of ($\mathrm{Z}^{(1)}$, $\mathrm{Z}^{(2)}$). For example, on institution A, there are m pairs of (${Z_{A1}}^{(1)}$, ${Z_{A1}}^{(2)}$), ⋯, (${Z_{Am}}^{(1)}$, ${Z_{Am}}^{(2)}$). And, after building ith model of institution A using ${Z_{Ai}}^{(1)}$, ith weight on the ith model of institution A is calculated by fitting the ith model of institution A into ${Z_{Ai}}^{(2)}$, ${Z_{Bi}}^{(2)}$ and ${Z_{Ci}}^{(2)}$. This process is also applied to other institution and m weights are generated for each institution. The better the model of the institution fits into overall data, the greater the weight on the model of the institution calculates well. The final weight on the model of each institution estimates as average of m weights and the integrated model is constructed based on estimated weight.

For an experiment, we used Fever Coach mobile healthcare application data for care of feverish children. We generated three horizontally distributed data which were extracted from vaccination data of Fever Coach collected between from September 2016 to December 2018. The sample size of each institution is 300 and we used logistic regression (LR) model as a model for validation.

Results

Among the variables in LR model, dependent variable is fever(body temperature≥39℃) for 72 hours after vaccination and independent variables are age(month), gender, febrile convulsion(yes or no), vaccination type (single or multiple vaccines) and duration time(hour, body temperature≥38℃ for 72 hours).

The predictive performances of LR models of the three institution were different, and the performance of institution A was better than that of the other two institutions. (Figure 1. (a)). We performed the weight-based LR model with 200 replications and final estimated weights were 0.2905, 0.3634 and 0.3461 on institution A, B, and C, respectively. The area under the curve(AUC) of the proposed LR model was 75.98%(72.89%-79.08%) and the centralized LR model was 75.72%(72.61%-78.83%) (Figure 1. (b)).

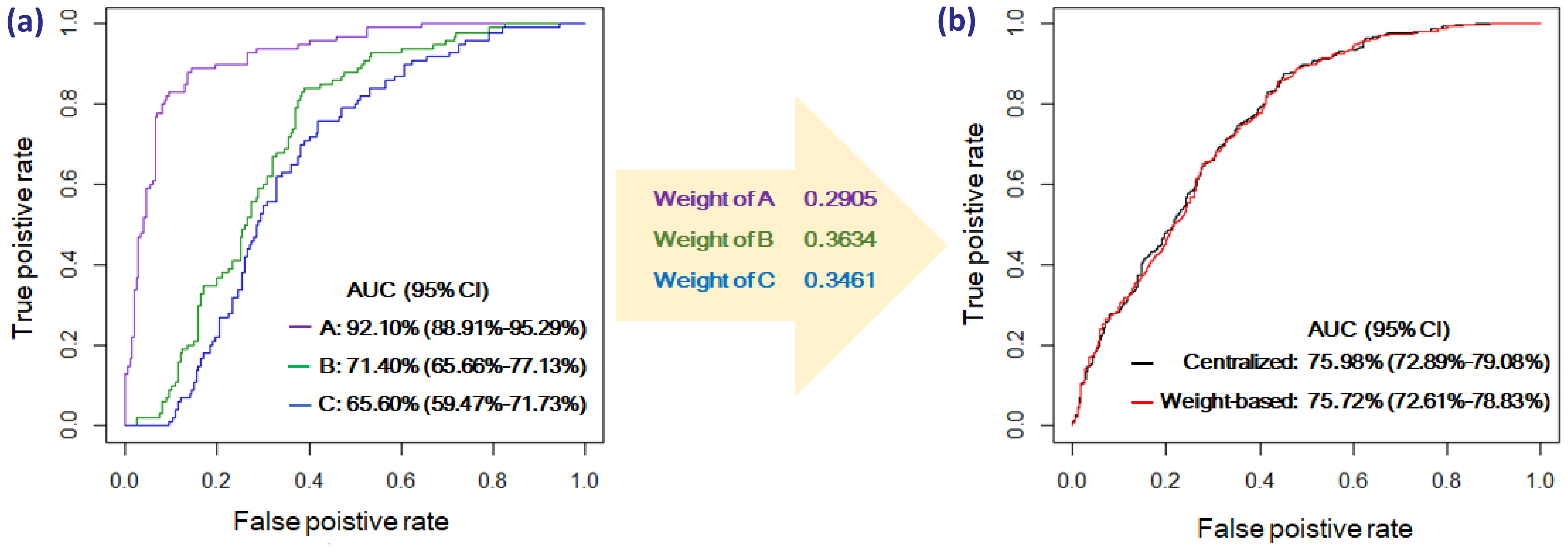

Figure 1. Results of weight-based integrated LR model.
(a) The ROC curve and AUC for the LR model for each Institution with size 300.
(b) The ROC curve and AUC fo the centralized LR model and weight-Based integratd LR model based on the total size 900.

This results show that the proposed model without sharing data had the same performance comparing with the centralized model with sharing data. The LR model of Institution A, with the smallest weight, has the largest differences in predictive value from centralized LR model than institution B and C for total data with size 900 (Figure 2).

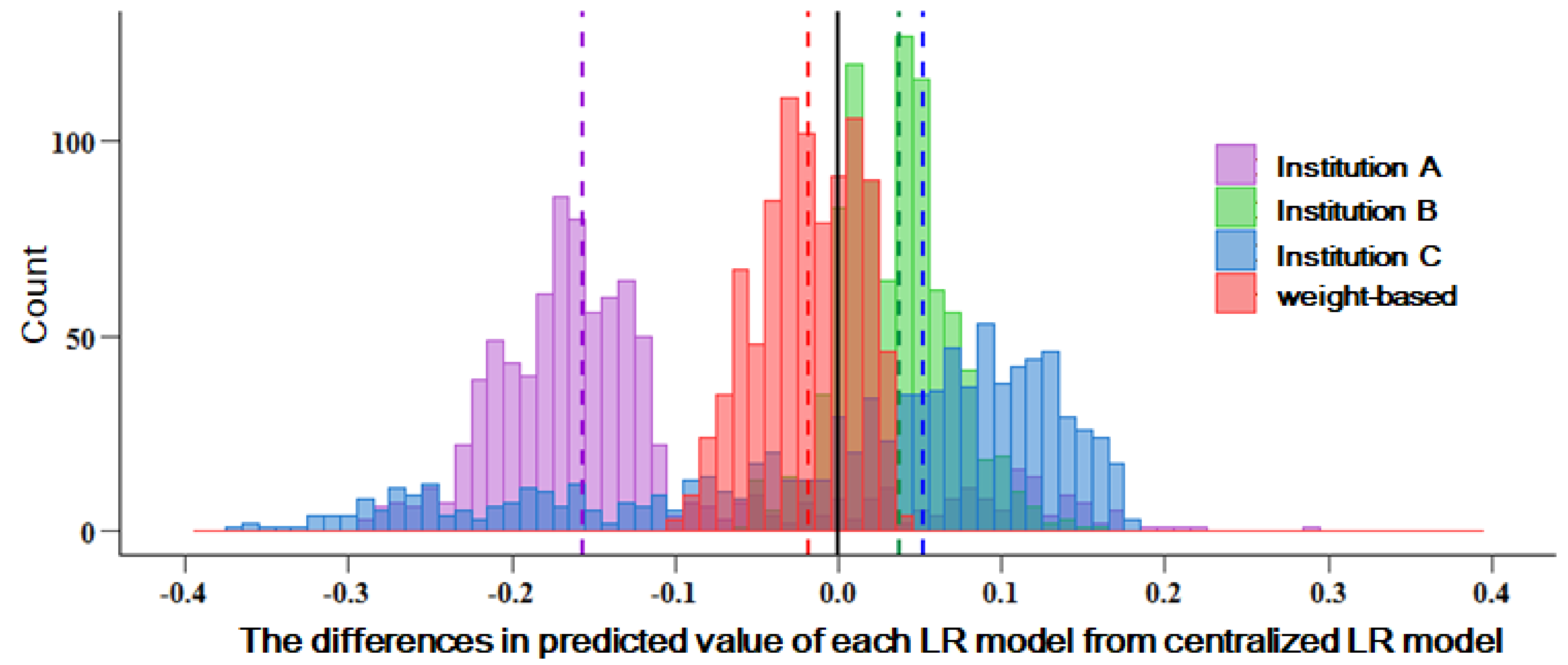

Figure 2. The differences in the predictive probability for all the 900 childrens, between each LR model and the centralized LR model. The vertical line presents median of the difference.

Conclusions

In this paper, It is shown that the weight-based LR model provides the same prediction as the centralized LR model through mobile application data-based simulation. The proposed model can help to build a prediction model that reflects the characteristics of various institutions in situations where data can't be shared.

References

[1] Jochems A, et al. Distributed learning: developing a predictive model based on data from multiple hospitals without data leaving the hospital–a real life proof of concept. Radiother. Oncol. 2016;121.3:459-467.
[2] El Emam K, et al. A secure protocol for protecting the identity of providers when disclosing data for disease surveillance. J. Am. Med. Inf. Assoc. 2011;18.3:212-217.

Acknowledgement

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : HI19C1234) and the Foundational Technology Development Program (NRF-2019M3E5D4064682) the Ministry of Science and ICT, Republic of Korea.