Performance of Logistic Regression in Artificial Intelligence for Public Health Issues: Meta-Analysis

Vineet Kumar Kamal[1]*and Dolly Kumari[2]

[1]Scientist D, ICMR-National Institute of Epidemiology, Chennai, India
[2]Assistant Professor, Center for Economic Policy and Public Finance, ADRI, Patna, India

Summary

Logistic regression and other artiýcial intelligence (AIs) models have great potential to accurately predict the outcome in the ýeld of public health. Also, logistic regression is the most popular choice as an analytic method if our mʿ t ɑ mʿ ʿ Th я mycem.
a meta-analysis study with the primary aim is to systematically review the performance of logistic regression used in public health and quantify those performances using meta-analysis, and secondary aim to compare the performance of logistic regression to other AIs or machine learning algorithms used in Public Health. A systematic literature search was performed to identify studies utilizing logistic regression in public health. The search strategy using the term ñlogistic regression [tiab] AND (machine learning [tiab] OR artiýcial intelligence [tiab]) AND public health [tiab]ò was performed with additional ýlters in PubMed. The performance of the models was quantiýed using area under curve with its 95% conýdence interval (AUC (95% CI)). A total of six studies were included after inclusion and exclusion criterion. The pooled AUC for logistic regression was 0.814 (95% CI 0.812 - 0.817) with high heterogeneity. Logistic regression, random forest, ANN, and gradient boosting model have similar performances with high discrimination ability (AUC > 0.80). It was concluded that logistic regression can be used as one of the ML algorithms, but random forest, ANN, and gradient bosting are also good performing models. However further research with more appropriate design is needed to conclude the same.

Keywords: Area under curve (AUC); Arti cial intelligence; Logistic regression; Machine learning; Public health

Introduction

Despite to be a newer area, arti cial intelligence (AI) / Machine Learning (ML) tools and techniques are advantageous in providing in depth knowledge on individuals' health and predicting population health risks, and their use for medicine as well as public health is expected to increase substantially in coming days. is eld is growing at an unprecedented pace in health care, including disease diagnosis, risk analysis, triage or screening, surgical operations, and so forth.

Logistic regression is one of the most popular supervised ML algorithms if our aim is to predict disease presence (diagnosis) or disease outcomes (prognosis) [1]. It is used for predicting the categorical response variable using a given set of independent/input/predictor variables. It examines the relationship between a binary outcome (dependent) variable such as presence or absence of disease and predictor (explanatory or independent) variables such as patient demographics or clinical ndings. For example, In-hospital mortality for head injury patients might be predicted from knowledge of the patient's age, Pupil reactivity, CT ndings and hypotension and other clinical parameters in Emergency department. e outcome variables can be both continuous and categorical. If X_1, X_2… X n denotes n predictor variables, Y denotes the hospital mortality (Y = 1), and p denotes the probability of hospital mortality (i.e., the probability that Y = 1), the following equation describes the relationship between the predictor variables and p:

$$\text{Log}\left(\frac{p}{14\,p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ... + \beta_n X_n,$$

Where, β_0 is a constant and β_2, β_2… β n are the regression coe cients of the predictor variables X_1, X_2… X_n. e regression coe cients are estimated from the available data and it represent the strength of the association between a patient characteristics and the outcome [2]. e machine learning over logistic regression for clinical prediction models [4]. It would be great beneficial for acceptance of logistic regression from modeling utility point of view if we can conclude the performance of logistic regression used in public health is high and can be used in alternative of other ML methods.

The aim of this study was to systematically review the performance of logistic regression in terms of AUC used in public health and quantify those performances using meta-analysis. Secondary aim was to compare the performance of logistic regression to other ML algorithms used in Public Health.

Materials and Methods

Literature Search Strategy

A systematic literature search was performed to identify studies utilizing logistic regression in public health. The search strategy using the term "logistic regression [tiab] AND (machine learning [tiab] OR artificial intelligence [tiab]) AND public health [tiab]" was performed in PubMed by author VKK. Then again search strategy was filtered using full text availability, publication in last one year, only English language, journal article and observational studies, limited to human.

Inclusion and exclusion criteria

Studies were eligible if

- The article used logistic regression as one of the ML models
- One of the performance measures was evaluated in terms of AUC with 95% CI

Studies were excluded

- The models made predictions for individual images or signals rather than participants
- Model based on tweet or social media

Prisma guideline

I followed the Preferred Reporting Items for Systematic reviews and Meta-Analysis (PRISMA) statement, though the study was registered with PROSPERO due to time constrain.

Statistical analysis

Data analysis was conducted as per guideline provided in the Handbook of Cochrane Systematic Review [5]. For the primary and secondary objectives, the effect measures (AUCs) were pooled along with their 95% confidence interval (CI) using the random-effects model with the Hedge's method considering substantial heterogeneity was present. Results were expressed through forest plot analysis and heterogeneity through Cochran's Q and I^2 statistic. All the analysis was performed using STATA 15.1.

Result

Description of Studies

The electronic search retrieved 78 articles. After applying filter, as given in method section, I ended with 27 articles. After putting inclusion and exclusion criterion, I found six articles. The description of author, sample size, methods, AUC with 95% CI, and outcome is described

95% CI for logistic regression ranged from 0.594 to 0.957 in all these six studies.

Results of Pooling

Logistic regression model

e pooled AUC for logistic regression was 0.814 (95% CI 0.812 - 0.817, I^2=99.8%) with high heterogeneity [Figure 1]. e test of θ, with z = 622.17, p = 0.000, indicates that the pooled results are statistically signi cant.

Random forest, ANN, and gradient boosting model

e pooled AUC for random forest model was 0.803 (95% CI 0.806 - 0.808, I^2=99.9%) with high heterogeneity [Figure 2]. e pooled AUC for ANN model was 0.824 (95% CI 0.822 - 0.827, I^2=99.9%) with high heterogeneity [Figure 3] [Figure 4]. Similarly, the pooled AUC for gradient boosting model was 0.828 (95% CI 0.826 - 0.831, I^2=99.9%). e test of θ all three models, with p = 0.000, indicates that the pooled results are statistically signi cant.

From [Table 2], it is clearly depicted that logistic regression, random forest, ANN, and gradient boosting model have high discrimination ability as AUC > 0.80, but all are performing more or less on same line, and I did not nd any statistically signi cant di erence among AUCs.

Discussion

We systematically reviewed performance of six arti cial intelligence-based or machine leering models used in public health for prediction. We found that logistic regression had high performance in terms of AUC. We also found that random forest, ANN, and gradient

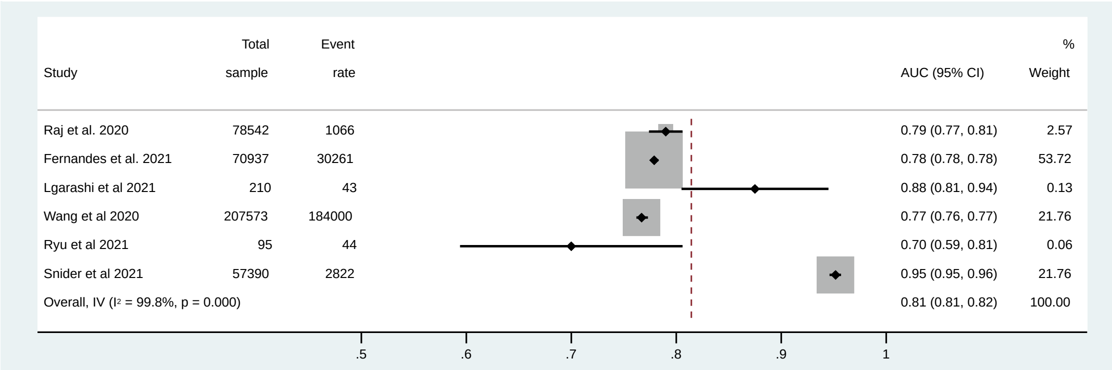
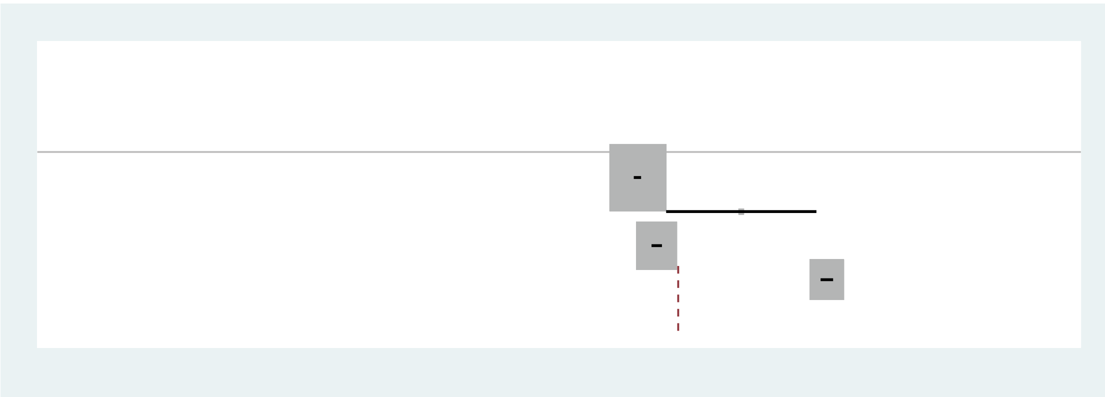


Figure 1: Forest plot for logistic regression model.



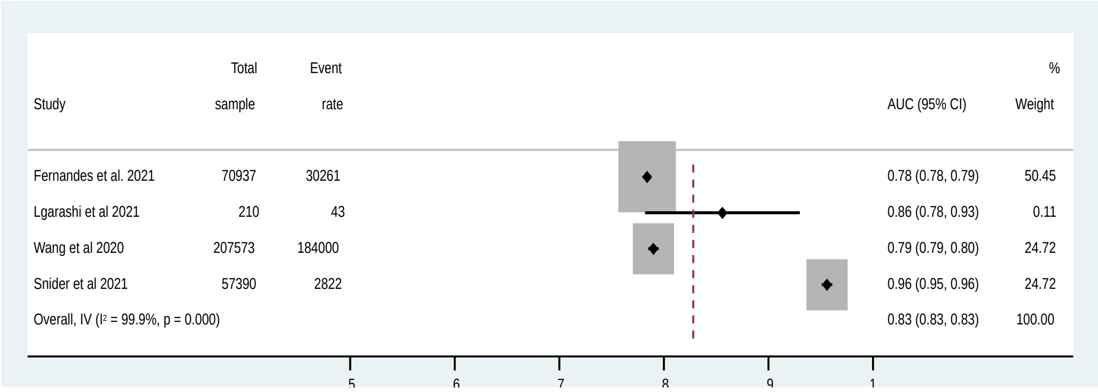


Figure 4: Forest plot for gradient booting model.