Comparison of machine-learning and logistic regression models to predict 30-day unplanned readmission: a development and validation study

Masao Iwagami PhD1,2,3, *, Ryota Inokuchi PhD1,2, *, Eiryo Kawakami PhD4,5, Tomohide Yamada PhD6, Atsushi Goto PhD7, Toshiki Kuno PhD8, Yohei Hashimoto PhD9,10, Nobuaki Michihata PhD11, Tadahiro Goto PhD10,12, Tomohiro Shinozaki PhD13, Yu Sun PhD1,2, Yuta Taniguchi MD1, Jun Komiyama MPH1,2, Kazuaki Uda PhD1,2, Toshikazu Abe PhD1,14, Nanako Tamiya PhD1,2

1Department of Health Services Research, Institute of Medicine, University of Tsukuba, Ibaraki, Japan
2Health Services Research and Development Center, University of Tsukuba, Ibaraki, Japan
3Faculty of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK
4Artificial Intelligence Medicine, Graduate School of Medicine, Chiba University, Chiba, Japan
5Advanced Data Science Project (ADSP), RIKEN, Yokohama, Japan
6Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
7Department of Health Data Science, Graduate School of Data Science, Yokohama City University, Kanagawa, Japan
8Division of Cardiology, Montefiore Medical Center, Albert Einstein College of Medicine, New York, USA
9Department of Ophthalmology, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
10Department of Clinical Epidemiology and Health Economics, School of Public Health, The University of Tokyo, Tokyo, Japan
11Department of Health Services Research, Graduate School of Medicine, The University of Tokyo, Tokyo, Japan
12TXP Medical Co. Ltd.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Department of Information and Computer Technology, Tokyo University of Science, Tokyo, Japan.

Department of Emergency and Critical Care Medicine, Tsukuba Memorial Hospital, Tsukuba, Japan.

*Contributed equally (co-first author)

*Corresponding author

**Corresponding author:**

Masao Iwagami, MD, MPH, MSc, PhD
Department of Health Services Research, Institute of Medicine, University of Tsukuba 1-1-1 Tennodai, Tsukuba, Ibaraki, Japan
Telephone: +81 (0)29 853 8849
Fax: +81 (0)29 853 8849
E-mail: iwagami-tky@umin.ac.jp
ORCID ID: 0000-0001-7079-0640

Full names and contact information of authors

Masao Iwagami, iwagami-tky@umin.ac.jp
Ryota Inokuchi, inokuchi.ryota.ge@u.tsukuba.ac.jp
Eiryo Kawakami, eiryo.kawakami@chiba-u.jp
Tomohide Yamada, bqx07367@yahoo.co.jp
Atsushi Goto, agoto@yokohama-cu.ac.jp
Toshiki Kuno, kuno-toshiki@hotmail.co.jp
Yohei Hashimoto, yohashimoto1223@gmail.com
Nobuaki Michihata, michihata@m.u-tokyo.ac.jp
Tadahiro Goto, tag695@mail.harvard.edu
Tomohiro Shinozaki, shinozaki@rs.tus.ac.jp
Funding: This study was supported by a Japan Society for the Promotion of Science (JSPS) KAKENHI Grant (No. 19K19430) from the Japanese Ministry of Education, Culture, Sports, Science, and Technology. The funder had no role in study design, data collection, data analysis, data interpretation, or writing.

Keywords: risk prediction, machine learning, artificial intelligence, readmission, electronic health records

Word count: 3,931

Number of tables: 1

Number of figures: 4

Number of supplementary tables: 8
Abstract (200/200 words)

We compared the predictive performance of gradient-boosted decision tree (GBDT), random forest (RF), deep neural network (DNN), and logistic regression (LR) with the least absolute shrinkage and selection operator (LASSO) for 30-day unplanned readmission, according to the number of predictor variables and presence/absence of blood-test results. We used electronic health records of patients discharged alive from 38 hospitals in 2015–2017 for derivation (n=339,513) and in 2018 for validation (n=118,074), including basic characteristics (age, sex, admission diagnosis category, number of hospitalizations in the past year, discharge location), diagnosis, surgery, procedure, and drug codes, and blood-test results. We created six patterns of datasets having different numbers of binary variables (that ≥5% or ≥1% of patients or ≥10 patients had) with and without blood-test results. For the dataset with the smallest number of variables (102), the c-statistic was highest for GBDT (0.740), followed by RF (0.734), LR-LASSO (0.720), and DNN (0.664). For the dataset with the largest number of variables (1543), the c-statistic was highest for GBDT (0.764), followed by LR-LASSO (0.755), RF (0.751), and DNN (0.720). We found that GBDT generally outperformed LR-LASSO, but the difference became smaller when the number of variables was increased and blood-test results were used.
Introduction

Unplanned hospital readmission is a common issue in public health because it imposes burdens on medical budgets, healthcare staff, and patients [1]. Some unplanned hospital readmission is preventable [2]. Therefore, it is important to identify patients with a high risk of unplanned readmission for intervention. For achieving this, a clinical prediction model that can accurately estimate the probabilities of clinical outcomes for individual patients is essential [3].

Clinical prediction models have been developed for 30- or 28-day unplanned hospital readmission—mostly logistic regression (LR) models based on a small number of clinical variables [4, 5]. However, the prediction ability of these models is often suboptimal, with a c-statistic or area under the receiver operating characteristic curve (AUROC) of <0.70. It was discussed that the prediction ability should be improved by increasing the number/variety of clinical variables and developing more accurate statistical or mathematical models with modern computing techniques [4, 5].

Recently, machine-learning models, such as the gradient-boosted decision tree (GBDT), random forest (RF), and deep neural network (DNN), have been applied to claims data and electronic health records (EHRs) to build prediction models for readmission [6-8]. However, according to a recent systematic review, there was no statistically significant difference in the predictive performance between “traditional regression models” (defined in the paper as LR or Cox regression models, regardless of feature selection techniques) and machine-learning models, with average c-statistic values of 0.71 for 26 studies and 0.74 for 15 studies, respectively, corresponding to a difference in c-statistic of 0.03 (95% confidence interval [CI]: –0.01 to 0.07) [8]. As its major limitation, however, only few studies included in the systematic review involved direct comparisons of “traditional regression models” and commonly used machine-learning models within the same study [9-13]. Moreover, the studies included in the systematic review tended to use a small number of predictor variables. The predictive performance of machine-learning models may depend on the number of predictor variables, as well as the inclusion of continuous variables potentially having nonlinear
relationships with a study outcome, such as blood-test results. Some machine-learning models may outperform an LR model if the number of predictor variables is increased and blood-test results are used for prediction.

Thus, using the EHRs obtained from 38 Japanese hospitals, we aimed to develop, validate, and compare GBDT, RF, DNN, and LR with the least absolute shrinkage and selection operator (LASSO) for predicting 30-day unplanned readmission, by intentionally creating six patterns of datasets having different numbers of predictor variables (which ≥5% or ≥1% of patients or ≥10 patients had) with and without blood-test results. Notably, among several feature-selection techniques for LR, such as stepwise variable selection and the LASSO, we decided to use the LASSO because it is generally less likely to cause overfitting than other techniques [14] and also because LR-LASSO outperformed LR with the stepwise variable selection in a study predicting 30-day all-cause non-elective readmission [11]. Our study hypothesis was that some machine-learning models (GBDT, RF, or DNN) outperform LR-LASSO more prominently if the number of predictor variables is increased and blood-test results are used. The findings of this study could help hospitals apply clinical prediction models to their EHRs to identify patients at high risk for readmission and intervene efficiently.

**Results**

In the Medical Data Vision (MDV) database, we identified 635,509 discharges of 410,941 patients, with at least one blood test during hospitalization, who were admitted after January 1, 2015 and discharged before December 31, 2018, from 38 hospitals (Figure 1). After the exclusion of admissions associated with childbirth, discharges dead, and transfers to other hospitals and cases of missing data, there were 457,587 discharges eligible for analysis, including 339,513 discharges (mean age 62.0 ± 24.6 years, 54.3% men) from January 1, 2015 to December 31, 2017 in the derivation dataset and 118,074 discharges (mean age 63.4 ± 24.1 years, 54.1% men) from January 1 to December
The incidence of 30-day unplanned readmission was 6.8% (23,108/339,513) and 6.4% (7,507/118,074) for the derivation and validation datasets, respectively. Table 1 presents the basic characteristics of the patients overall and by outcome status. Patients with 30-day unplanned readmission tended to be older men and were more likely to have been hospitalized in the past year, admitted for diseases of the respiratory or digestive system, and discharged to nursing homes than those without 30-day unplanned readmission.

Among the 2102 International Classification Disease 10th revision (ICD-10) codes, 914 surgery codes, 122 procedure codes, and 468 Anatomical Therapeutic Chemical (ATC) codes, we excluded variables that <10 patients had, in both the derivation and validation datasets, to avoid the “perfect separation” (or “complete separation”) problem [15]. Consequently, 823 ICD-10 codes (Supplementary Table S1), 259 surgery codes (Supplementary Table S2), 65 procedure codes (Supplementary Table S3), and 381 ATC codes (Supplementary Table S4) were regarded as candidate binary predictor variables for the following analyses. The distribution of the 10 types of blood-test results, as the last measurement during hospitalization, is presented in Supplementary Table S5.

Then, we intentionally created six patterns of datasets with different numbers of the candidate binary predictor variables (that ≥5% or ≥1% of patients or ≥10 patients had, in both the derivation and validation datasets) with and without blood-test results, in addition to the basic characteristics presented in Table 1. The number of candidate predictor variables was 102 for pattern 1 (including binary variables that ≥5% had, without blood-test results), 112 for pattern 2 (including binary variables that ≥5% had, with blood-test results), 296 for pattern 3 (including binary variables that ≥1% had, without blood-test results), 306 for pattern 4 (including binary variables that ≥1% had, with blood-test results), 1533 for pattern 5 (including binary variables that ≥10 patients had, without blood-test results), and 1543 for pattern 6 (including binary variables that ≥10 patients had, with
For each pattern, using the derivation dataset, each model was developed with optimized hyperparameters (Supplementary Table S6). By applying these models to the validation dataset, the discrimination ability of each model—indicated by the c-statistic or AUROC—was evaluated, as shown in Figure 2 and Supplementary Table S7. The GBDT outperformed LR-LASSO for all patterns (Delong’s test P-value <0.001). The discrimination ability of DNN was the worst for all patterns. In more detail, for pattern 1 (including binary variables that ≥5% had, without blood-test results), which had the smallest number of variables, the point estimate in c-statistic was the highest for GBDT (0.740), followed by RF (0.734), LR-LASSO (0.720), and DNN (0.664). For pattern 6 (including binary variables that ≥10 patients had, with blood-test results), which had the largest number of variables, the point estimate in c-statistic was highest for GBDT (0.764), followed by LR-LASSO (0.755), RF (0.751), and DNN (0.720), suggesting that the difference between GBDT and LR-LASSO was small.

For pattern 6 (including binary variables that ≥10 patients had, with blood-test results), the ROC curves are shown in Figure 3, and the calibration plots are shown in Figure 4. As graphically shown, the calibration of every model was similarly good for patients at low risk for readmission. However, GBDT and LR-LASSO tended to overestimate, whereas RF and DNN tended to underestimate the probability of readmission for patients at higher risk for readmission.

For pattern 6 (including binary variables that ≥10 patients had, with blood-test results), Supplementary Table S8 presents all regression coefficients in LR-LASSO, whereas Figure 5 shows the variable importance of the top 10 important predictors for GBDT, suggesting that age, blood-test results, and number of hospitalizations in the past year were important predictors.

Discussion

Using the EHRs of 38 Japanese hospitals, we systematically compared the prediction performance for
30-day unplanned readmission among commonly used machine-learning models (GBDT, RF, and DNN) and an LR-LASSO model, by intentionally creating six patterns of datasets having different numbers of predictor variables (that $\geq 5\%$ or $\geq 1\%$ of patients or $\geq 10$ patients had) with and without blood-test results. The discrimination ability mostly improved if the number of predictor variables was increased or blood-test results were added. For the latter patterns of datasets, the c-statistics of the models except for DNN were approximately 0.75 or higher, suggesting that the performance of these models was better than or comparable to that of previously reported models [4, 5, 8]. For each pattern of dataset, the c-statistic of GBDT was higher than that of LR-LASSO. However, against our hypothesis (that some machine-learning models outperform LR-LASSO more prominently if the number of predictor variables is increased and blood-test results are used), the difference in c-statistics between GBDT and LR-LASSO became rather smaller in the last pattern of dataset, which had the largest number of binary variables and blood-test results.

To our knowledge, this was the first systematic comparison of the predictive performance of commonly used machine-learning models and an LR model for unplanned readmission, using datasets characterized by different numbers of predictor variables and the presence/absence of blood-test results. A recent systematic review [8] included several studies comparing the performance of different prediction models for readmission with the same dataset; however, the target population was narrow, e.g., patients with lupus [12] and heart failure [13]. Our study is partly similar to that of Jamei et al. [9], who used the EHRs of $>300,000$ patients in Northern California, USA to compare the predictive performance of RF with 100 features; NN with 100, 500, and 1667 features; and LR with 100, 500, and 1667 features. With 100 features, RF and NN outperformed LR, with AUROCs of 0.77, 0.76, and 0.72, respectively. With 1667 features, NN performed far better than LR, with AUROCs of 0.78 and 0.66, respectively. We speculate that the reduction in the predictive performance of LR according to an increase in the number of features was probably due to overfitting, as it appears that the authors did not conduct any regularization or penalization via means such as LASSO, which we
In the present study, the predictive performance of DNN was worse than that of the other models. This is probably because our dataset, which included mostly binary variables and only a small number of continuous variables (such as age and blood-test results), was simpler than the dataset of Jamei et al., which had a larger number of continuous variables, such as vital signs and blood-test results.

To date, comparisons of machine-learning models and “traditional regression models” (defined as LR or Cox regression models, regardless of feature selection techniques) have been performed for various outcomes. Christodoulou et al. conducted a systematic review of 71 studies on clinical prediction models, regardless of outcomes, and concluded that there was no evidence of machine-learning methods outperforming LR. However, the systematic review was limited in that it included only studies with <100 predictor variables; the authors identified seven studies with >100 predictor variables, all of which were excluded owing to “high risk of bias” in their classification. Therefore, it remained unclear whether machine-learning methods can outperform LR if the dataset includes a larger number of predictor variables. Our study answers this question: we intentionally created six patterns of datasets having different numbers of predictor variables (102, 112, 296, 306, 1533, or 1543) with and without blood-test results, and in the dataset with the largest number of variables, the difference in c-statistics between GBDT and LR was rather small. Therefore, the number of predictor variables does not seem to affect the results and conclusions very much when comparing the predictive performance of the machine-learning and LR models.

Instead, the types or characteristics of predictors may matter. Notably, in the present study, the vast majority of the predictor variables were binary, and the number of continuous variables was small. An increased number of continuous variables, particularly those having nonlinear relationships with a study outcome, and the use of more complex information (such as images and sounds) may be needed to demonstrate the superiority of machine-learning methods over “traditional regression models” in clinical risk prediction. Further studies are needed to clarify this aspect.
Regarding clinical implications, according to the results of the present study, for hospitals applying clinical prediction models to their EHRs to identify patients at high risk for readmission, the GBDT may be recommended. However, the benefit of GBDT over LR-LASSO was small in the present study, suggesting that “traditional regression models” such as an LR model may also be acceptable, considering that they are more user-friendly and easier to interpret. In GBDT, the variable importance was high for age, blood-test results, number of hospitalizations in the previous year. Therefore, clinicians should be conscious of these variables when their patients are to be discharged. Patients at high risk for readmission may need to have their discharge postponed or receive intensive post-discharge follow-up as an outpatient [2].

This study had several limitations. First, we used data from a relatively large number of hospitals, but it is unclear whether these hospitals are representative of Japan. Although we believe that the selection of these hospitals is unlikely to be associated with the performance of the machine-learning models, our findings should be validated using data from different hospitals in Japan as well as hospitals in other countries. Second, we were unable to differentiate between patients admitted to different hospitals, because the hospital identifier was omitted to preserve privacy and data safety in the MDV database. Therefore, we could not account for any clustering effect by hospitals in our analysis. Third, the incidence of 30-day unplanned readmission was meaningfully lower in 2018 in the validation dataset (6.4%) than that in 2015-2017 in the derivation dataset (6.8%), possibly suggesting a decreasing trend in hospital readmission in Japan. This trend might have harmed the applicability of the developed models to the validation dataset, resulting in underestimation of the predictive ability of the models. Fourth, although the MDV database contained important information on recorded diagnoses and treatments during hospitalization, as well as blood-test results, it lacks other clinical variables that may be useful for the prediction of hospital readmission, such as vital signs and socioeconomic features [8]. Finally, we used a hospital-based database and could only predict readmission to the hospital that the patient was initially admitted to, rather than readmission to
any hospital. In emergencies, patients may be sent to another hospital for readmission. Thus, a study utilizing a population-based database should be conducted to assess the performance of machine-learning methods in predicting readmission to any hospital.

**Conclusions**

Using the EHRs of 38 hospitals in the MDV database, we compared the predictive performance of machine-learning and LR models for six patterns of datasets having different numbers of predictor variables with and without blood-test results. For every pattern, GBDT outperformed LR-LASSO. However, the difference in c-statistic between GBDT and LR-LASSO was small, especially in the dataset with the largest number of variables (1543 variables including blood-test results). Thus, instead of the number of predictor variables, future studies should focus on the types or characteristics of predictors in EHRs for demonstrating the superior performance of machine-learning methods to “traditional regression models”.

**Methods**

**Design and setting**

We conducted a retrospective cohort study using a hospital-based database in Japan.

**Data source**

Diagnosis Procedure Combination (DPC) data in Japan are administrative claims data and discharge summaries, which are collected at the top of the DPC system, a case-mix patient classification, and a lump-sum payment system for inpatients in acute care hospitals [17]. The MDV database, which was built by Medical Data Vision Co., Ltd. (Tokyo, Japan), consisted of >350 acute care hospitals—approximately 20% of the DPC hospitals in Japan. It consisted of hospitals that used the business support system of MDV Co., Ltd. and provided consent for secondary data use for research.
purposes. The age distribution of inpatients in the MDV database is similar to that of all DPC hospitals, whereas the hospital volume (i.e., number of beds) tends to be larger [18]. During the study period between January 1, 2015 and January 31, 2019, laboratory test values were available from 38 hospitals that agreed to provide the data. Thus, in this study, we used DPC data (administrative claims data and discharge summaries) and blood-test results from 38 hospitals. We did not conduct any sample size calculation, but planned to use all the available data.

The DPC data included basic patient characteristics such as age and sex; recorded diagnoses based on ICD-10 codes [19], including the admission diagnosis (indicating the reason for hospital admission), main diagnosis, most resource-consuming diagnosis, second-most resource-consuming diagnosis, comorbidities present on admission (up to four diagnoses), and complications arising after admission (up to four diagnoses), recorded by the responsible physician at the time of discharge; Japanese original surgery codes [20], if a patient received surgery during hospitalization; Japanese original procedure codes [21], such as mechanical ventilation and dialysis; and original Japanese prescription codes (including both oral and intravenous drugs), which are linked to the ATC codes [22]. In addition, the admission status (planned or unplanned) and discharge status (dead, discharged to nursing facility or home, transferred to other hospitals) are available.

The study was approved by the Ethics Committee of the University of Tsukuba (approval no. 1414) in accordance with the Declaration of Helsinki. Because the claims data were anonymized before the researchers received them, individual participants’ consent was waived according to the ethical guidelines for medical and health research involving human subjects [23]. To ensure privacy and data safety, the hospital identifier was omitted by MDV Co., Ltd. in the data that we received. Thus, we could not identify the names and characteristics of the 38 hospitals and did not know which of the 38 hospitals the patients were admitted to.

Study participants
We identified the discharges of patients who were admitted to the 38 hospitals, with at least one blood-test during hospitalization, from January 1, 2015 to December 31, 2018. We excluded: i) cases with the admission diagnosis category suggesting childbirth (ICD-10 O00-Q99), ii) discharges dead, iii) transfers to other hospitals, and iv) cases with missing values of discharge location, or one or more of 10 basic blood-tests (white cell count, hemoglobin, platelet count, sodium, potassium, chloride, creatinine, blood urea nitrogen, aspartate transaminase, and alanine transaminase). We excluded cases with missing values of one or more of 10 basic blood-tests because (i) missingness is not likely to be at random, and (ii) even if missingness is at random, strategies to impute missing data in different machine-learning models are different or not established [24], whereas the objective of the present study was to compare the predictive performance of different models for the same complete dataset. If the same patient was hospitalized multiple times during the study period, we considered each admission as an independent admission, while we used the number of hospitalizations in the past year as a predictor variable to reduce dependence between them.

The data were split into derivation data for patients discharged from January 1, 2015 to December 31, 2017, which were used to develop the models, and validation data for patients discharged from January 1 to December 31, 2018, which were used to assess the discrimination ability and calibration of the models.

**Outcome**

The outcome of interest was 30-day unplanned readmission to the hospital from which the patient was initially discharged.

**Predictor variables and datasets**

The predictor variables used in this study consisted of basic characteristics, recorded diagnoses, inpatient treatments, and blood-test results that were the last measurement before discharge. The basic
characteristics included age (as a continuous variable), sex, admission diagnosis category (according to the ICD-10 codes, from A to T), number of hospitalizations in the past year, and discharge location (home or nursing home). For recorded diagnoses, information on ICD-10 codes recorded anywhere in the DPC code position (i.e., admission, main, most or second-most resource-consuming, comorbidities present on admission, or complications arising after admission) was transformed into the presence or absence of each 3-digit ICD-10 code from A00 to Z00. In total, 2102 types of 3-digit ICD-10 codes in the WHO ICD-10 Version 2010 were used in this study [19]. Similarly, for inpatient treatments, information on the surgery, procedure, and prescribed drugs recorded in the DPC system was transformed into the presence or absence of 914 surgery codes from K000 to K939 [20], 122 procedure codes from J000 to J129 [21], and 468 ATC codes from A01A0 to V07A0 [22].

Among the 2102 ICD-10 codes, 914 surgery codes, 122 procedure codes, and 468 ATC codes, we excluded variables that <10 patients had, in both the derivation and validation datasets, to avoid the “perfect separation” (or “complete separation”) problem [15]. Then, using these candidate binary predictor variables, we created six patterns of datasets having different numbers of the variables (that $\geq 5\%$ or $\geq 1\%$ of patients or $\geq 10$ patients, in both the derivation and validation datasets) with and without blood-test results, in addition to the basic characteristics (i.e., age, sex, admission diagnosis category, number of hospitalizations in the past year, and discharge location).

The continuous variables, including age, the number of hospitalizations in the past year, and blood test results, were standardized to a mean of 0 and a standard deviation of 1. The categorical variables, including the number of hospitalizations in the past year and admission diagnosis category (16 values from A to T), were transformed into a combination of binary variables using dummy values.

**Machine-learning and LR models**

Details regarding the supervised machine-learning models are presented elsewhere [15, 25]. In brief,
RF comprises ensembles of decision trees constructed from bootstrapped training samples, for which random samples corresponding to specific numbers of predictors are selected to initiate tree induction. The GBDT is another ensemble method, in which new tree models for predicting the errors and residuals of previous models are constructed in sequence. These new models are combined, and a gradient descent algorithm is used to minimize the loss function. A DNN comprises multiple processing layers and model outcomes via intermediate hidden units, each comprising a linear combination of predictors that are transformed into nonlinear functions.

There are several feature-selection techniques for LR, such as stepwise variable selection and the LASSO [3]. In this study, we used the LASSO because it is generally less likely to cause overfitting than other techniques [14]. Additionally, in a previous study, LR with the LASSO outperformed stepwise variable selection in predicting 30-day all-cause non-elective readmission [11]. We did not include any interaction terms between the features in our LR-LASSO.

**Statistical analysis**

First, we described the distribution or proportions of all the predictor variables according to outcome status in the derivation and validation data. Next, for each pattern of datasets from 1 to 6, we used derivation data to establish the machine-learning and LR models. The hyperparameters for each model were determined and optimized via 10-fold cross-validation within the derivation data (i.e., training data) [26], using automated machine learning in the h2o package of R. Then, using the validation data (i.e., test data), we evaluated the performance of each model with the c-statistic or AUROC, as well as calibration plots. We conducted Delong’s tests to compare the c-statistics of machine-learning models and LR-LASSO as a reference.

By focusing on GBDT, which showed the best performance among the studied models, we identified the 10 variables with the highest importance to examine the contribution of each predictor to the model with the best discriminatory abilities [27].
Data cleaning was conducted using STATA version 16 (StatCorp LLC, Texas, USA), and a statistical analysis was performed using R version 4.1.2 (R Foundation for Statistical Computing, Vienna, Austria) with the h2o and rms packages.

We followed the checklist of the Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD) Statement [28].

Acknowledgements

We express our gratitude to the personnel of Medical Data Vision Co., Ltd.—particularly Masaki Nakamura, Shogo Atsuzawa, and Masayoshi Suzuki—for their contributions to the preparation of the data. We thank Rina Yamauchi at the School of Public Health, University of Tokyo Graduate School of Medicine for her assistance in creating Supplementary Tables from S1 to S4. We also thank Editage (www.editage.com) for the English language editing.

Author contributions

M.I. planned the study, obtained funding and data, and wrote the main manuscript text. M.I. and R.I. analyzed the data and prepared all the tables and figures. N.M., Y.T., J.K., and K.U. assisted with data cleaning. E.K., T.Y., A.G., T.K., Y.H., T.S., Y.S., and T.A. contributed substantially to the analysis using machine-learning models and interpretation of the results. N.T. supervised the study as the laboratory head. All authors reviewed the manuscript and provided substantial feedback.

Data sharing

We obtained data from Medical Data Vision Co., Ltd. (MDV) and are not allowed to share these data with other parties. Researchers who meet the criteria for access can acquire de-identified participant data from MDV (https://en.mdv.co.jp).
Declaration of interests

We have no conflicts of interest to declare.

References


Figure legends

Figure 1. Flowchart
DPC, Diagnosis Procedure Combination; MDV, Medical Data Vision.

Figure 2. C-statistic of each model for six patterns of datasets having different numbers of the
candidate predictor variables with and without blood-test results

AUROC, area under the receiver operating characteristic curve; CI, confidence interval; GBDT, gradient-boosted decision tree; RF, random forest; DNN, deep neural network; LR-LASSO, logistic regression with the least absolute shrinkage and selection operator.

Pattern 1: Basic characteristics and binary predictor variables that \( \geq 5\% \) of patients had, without blood-test results (102 variables)

Pattern 2: Basic characteristics and binary predictor variables that \( \geq 5\% \) of patients had, without blood-test results (112 variables)

Pattern 3: Basic characteristics and binary predictor variables that \( \geq 1\% \) of patients had, without blood-test results (296 variables)

Pattern 4: Basic characteristics and binary predictor variables that \( \geq 1\% \) of patients had, with blood-test results (306 variables)

Pattern 5: Basic characteristics and binary predictor variables that \( \geq 10 \) patients had, without blood-test results (1533 variables)

Pattern 6: Basic characteristics and binary predictor variables that \( \geq 10 \) patients had, with blood-test results (1543 variables)

Figure 3. Receiver operating characteristic curve of each model for the dataset with the largest number of variables including blood-test results (1543 variables)

GBDT, gradient-boosted decision tree; RF, random forest; DNN, deep neural network; LR-LASSO, logistic regression with the least absolute shrinkage and selection operator.

Figure 4. Calibration plot of each model for the dataset with the largest number of variables including blood-test results (1543 variables)

GBDT, gradient-boosted decision tree; RF, random forest; DNN, deep neural network; LR-LASSO, logistic regression with the least absolute shrinkage and selection operator.
Figure 5. Variable importance of the top 10 important predictors in gradient-boosted decision tree for the dataset with the largest number of variables including blood-test results (1543 variables).

ICD-10, International Classification Disease 10th revision; BUN, blood urea nitrogen; WBC, white blood cell.

List of supplementary materials

Supplementary Table S1. Details and frequency of diagnosis codes (International Classification of Diseases, 10th revision)
Supplementary Table S2. Details and frequency of surgery codes (Japanese original codes)
Supplementary Table S3. Details and frequency of procedure codes (Japanese original codes)
Supplementary Table S4. Details and frequency of drug codes (Anatomical Therapeutic Chemical codes)
Supplementary Table S5. Details and distribution of blood-test results
Supplementary Table S6. Hyperparameters in each model for the dataset with the largest number of variables including blood-test results (1543 variables)
Supplementary Table S7. Discrimination ability (c-statistic) of each model for validation
Supplementary Table S8. Regression coefficients in logistic regression with the least absolute shrinkage and selection operator for the dataset with the largest number of variables including blood-test results (1543 variables)
Table 1. Baseline characteristics of patients

<table>
<thead>
<tr>
<th></th>
<th>Derivation dataset (N=339,513)</th>
<th>Validation dataset (N=118,074)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Outcome (-)</td>
</tr>
<tr>
<td>Age, mean±SD</td>
<td>62.0±24.6</td>
<td>61.5±24.7</td>
</tr>
<tr>
<td>Sex (men), n (%)</td>
<td>184,218 (54.3)</td>
<td>170,962 (54.0)</td>
</tr>
<tr>
<td>No. of hospitalization in the past year, n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>209,597 (61.7)</td>
<td>200,083 (63.2)</td>
</tr>
<tr>
<td>1</td>
<td>65,259 (19.2)</td>
<td>60,075 (19.0)</td>
</tr>
<tr>
<td>≥2</td>
<td>64,657 (19.0)</td>
<td>56,247 (17.8)</td>
</tr>
<tr>
<td>Admission diagnosis category (ICD-10 code), n (%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Certain infectious and parasitic diseases (A00-B99)</td>
<td>10,555 (3.1)</td>
<td>9,865 (3.1)</td>
</tr>
<tr>
<td>Neoplasms (C00-D48)</td>
<td>92,994 (27.4)</td>
<td>86,579 (27.4)</td>
</tr>
<tr>
<td>Diseases of the blood, blood-forming organs and immune mechanism (D50-D89)</td>
<td>3,525 (1.0)</td>
<td>3,086 (1.0)</td>
</tr>
<tr>
<td>Endocrine, nutritional and metabolic diseases (E00-E90)</td>
<td>11,117 (3.3)</td>
<td>10,309 (3.3)</td>
</tr>
<tr>
<td>Mental and behavioural disorders (F00-F99)</td>
<td>402 (0.1)</td>
<td>382 (0.1)</td>
</tr>
<tr>
<td>Diseases of the nervous system (G00-G99)</td>
<td>6,153 (1.8)</td>
<td>5,811 (1.8)</td>
</tr>
<tr>
<td>Diseases of the eye and adnexa (H00-H59)</td>
<td>741 (0.2)</td>
<td>709 (0.2)</td>
</tr>
<tr>
<td>Diseases of the ear and mastoid process (H60-H95)</td>
<td>3,481 (1.0)</td>
<td>3,533 (1.1)</td>
</tr>
<tr>
<td>Diseases of the circulatory system (I00-I99)</td>
<td>48,061 (14.2)</td>
<td>45,330 (14.3)</td>
</tr>
<tr>
<td>Diseases of the respiratory system (J00-J99)</td>
<td>45,224 (13.3)</td>
<td>41,191 (13.0)</td>
</tr>
<tr>
<td>Diseases of the digestive system (K00-K93)</td>
<td>47,648 (14.0)</td>
<td>43,846 (13.9)</td>
</tr>
<tr>
<td>Diseases of the skin and subcutaneous tissue (L00-L99)</td>
<td>4,233 (1.2)</td>
<td>3,954 (1.2)</td>
</tr>
<tr>
<td>Diseases of the musculoskeletal system and connective tissue (M00-M99)</td>
<td>13,387 (3.9)</td>
<td>12,945 (4.1)</td>
</tr>
<tr>
<td>Diseases of the genitourinary system (N00-N99)</td>
<td>19,440 (5.7)</td>
<td>18,078 (5.7)</td>
</tr>
<tr>
<td>Symptoms, signs and abnormal clinical and laboratory findings (R00-R99)</td>
<td>6,516 (1.9)</td>
<td>5,982 (1.9)</td>
</tr>
<tr>
<td>Injury, poisoning and certain other consequences of external causes (S00-T98)</td>
<td>26,036 (7.7)</td>
<td>24,985 (7.9)</td>
</tr>
<tr>
<td>Discharge place (discharge to nursing home [vs. home]), n (%)</td>
<td>19,214 (5.7)</td>
<td>16,917 (5.3)</td>
</tr>
</tbody>
</table>

SD = standard deviation, ICD-10 = International Classification of Diseases 10th Revision.
Any hospital discharge in 38 DPC hospitals with at least one blood-test between January 1, 2015 and December 31, 2018 in the MVD database. N = 635,509 discharges (410,941 patients).

- Admission for childbirth: n=21,317
- Discharge dead: n=26,931
- Transfer to other hospitals: n=56,424
- Missing data: n=78,378
  - Missing information on destination: n=849
  - Missing ≥1 blood-test variable: n=72,401

Analysis dataset: n=457,587

**Derivation dataset**
Discharge between January 1, 2015 and December 31, 2017
n=339,513

- 30-day unplanned readmission n = 23,108 (6.8%)

**Validation dataset**
Discharge between January 1 and December 31, 2018
n=118,074

- 30-day unplanned readmission n = 7,507 (6.4%)