PLOS ONE

RESEARCH ARTICLE

Claims-based algorithms for common chronic conditions were efficiently constructed using machine learning methods

Konan Hara[1], Yasuki Kobayashi[1], Jun Tomio[1], Yuki Ito[2], Thomas Svensson[3,4,5], Ryo Ikesu[1,3], Ung-il Chung[3,5,6], Akiko Kishi Svensson[3,4,7]*

1 Department of Public Health, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan, **2** Department of Economics, University of California, Berkeley, Berkeley, California, United States of America, **3** Precision Health, Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, Bunkyo-ku, Tokyo, Japan, **4** Department of Clinical Sciences, Lund University, Skåne University Hospital, Malmö, Sweden, **5** School of Health Innovation, Kanagawa University of Human Services, Kawasaki-shi, Kanagawa, Japan, **6** Clinical Biotechnology, Center for Disease Biology and Integrative Medicine, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan, **7** Department of Diabetes and Metabolic Diseases, Graduate School of Medicine, The University of Tokyo, Bunkyo-ku, Tokyo, Japan

* kishi@bioeng.t.u-tokyo.ac.jp

Abstract

Identification of medical conditions using claims data is generally conducted with algorithms based on subject-matter knowledge. However, these claims-based algorithms (CBAs) are highly dependent on the knowledge level and not necessarily optimized for target conditions. We investigated whether machine learning methods can supplement researchers' knowledge of target conditions in building CBAs. Retrospective cohort study using a claims database combined with annual health check-up results of employees' health insurance programs for fiscal year 2016–17 in Japan (study population for hypertension, N = 631,289; diabetes, N = 152,368; dyslipidemia, N = 614,434). We constructed CBAs with logistic regression, k-nearest neighbor, support vector machine, penalized logistic regression, tree-based model, and neural network for identifying patients with three common chronic conditions: hypertension, diabetes, and dyslipidemia. We then compared their association measures using a completely hold-out test set (25% of the study population). Among the test cohorts of 157,822, 38,092, and 153,608 enrollees for hypertension, diabetes, and dyslipidemia, 25.4%, 8.4%, and 38.7% of them had a diagnosis of the corresponding condition. The areas under the receiver operating characteristic curve (AUCs) of the logistic regression with/without subject-matter knowledge about the target condition were .923/.921 for hypertension, .957/.938 for diabetes, and .739/.747 for dyslipidemia. The logistic lasso, logistic elastic-net, and tree-based methods yielded AUCs comparable to those of the logistic regression with subject-matter knowledge: .923-.931 for hypertension; .958-.966 for diabetes; .747-.773 for dyslipidemia. We found that machine learning methods can attain AUCs comparable to the conventional knowledge-based method in building CBAs.

OPEN ACCESS

Citation: Hara K, Kobayashi Y, Tomio J, Ito Y, Svensson T, Ikesu R, et al. (2021) Claims-based algorithms for common chronic conditions were efficiently constructed using machine learning methods. PLoS ONE 16(9): e0254394. https://doi.org/10.1371/journal.pone.0254394

Editor: Chi-Hua Chen, Fuzhou University, CHINA

Received: February 11, 2021

Accepted: June 25, 2021

Published: September 27, 2021

Data Availability Statement: We cannot publicly provide individual data due to participant privacy, according to ethical guidelines in Japan. Qualifying researchers may apply to access a minimal dataset by contacting Precision Health, Department of Bioengineering, Graduate School of Engineering, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan, bpm@coi.t.u-tokyo.ac.jp. R codes are available at https://github.com/harakonan/research-public/tree/master/cba.

Funding: This research is supported by the Center of Innovation Program from Japan Science and

Technology Agency, JST. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

A growing body of studies using medical and pharmacy claims data has been conducted in various fields of health research [1–7]. Among them, a notable amount of research has used claims data to assess medical conditions [1, 2, 6]. Despite its large volume of information and highly standardized format, however, claims data is frequently criticized for its potential imprecision in the identification of medical conditions mainly because they are primarily issued for reimbursement purpose [8–12].

To address these concerns, plenty of studies have proposed a claims-based algorithm (CBA) for identifying patients with their target condition and computed association measures to assess the usability of the algorithm [9, 10, 13–41]. Previous studies have engaged in a knowledge-based condition-specific CBA construction procedure, i.e., researchers selected input variables and decided how to incorporate them in the CBA based on their experience or existing clinical knowledge regarding the target condition. Although this approach is widely used and intuitively plausible, it is highly dependent on the level of knowledge on the target conditions and is hard to obtain appropriate and reproducible CBAs. This is imposing challenges to the use of administrative data in the transition from the ICD-9 to the ICD-10 coding scheme in the United States [42, 43].

Moreover, since previous CBA studies are predominantly coming from North American countries, research using diagnosis derived from North American countries' claims data can be largely backed by a corresponding CBA study. In contrast, despite the rapid increase of research using diagnosis derived from claims data in other countries–e.g., Japan and Taiwan–CBAs are not established for most medical conditions thus far [44]. It is notable that the lack of confirmed CBA not only degrades the quality of research but also makes the research extremely difficult to be accepted by journals with high impact factors [45]. For this reason, researchers who are using claims data in these countries are facing an urgent need to establish CBAs for various medical conditions.

To this end, some researchers applied conventional regression methods to develop CBAs which are less dependent on the knowledge [9, 14, 17, 19, 24, 35, 37, 40]. However, the selection of input variables is required before implementing a regression model to obtain a satisfactory CBA, as conventional regression methods often work poorly in prediction accuracy when the number of input variables is large relative to the sample size [46]. Besides, if researchers expect nonlinear or interactive effects of the input variables, they have to specify those terms *a priori* as a functional form of the regression model.

Machine learning methods are promising technologies to overcome the problems of conventional regression methods, and some researchers have attempted to use these methods in the context of CBA [18, 25, 29, 30, 39]. However, they selected the input variables according to their target condition. Thus, to apply their procedures to other conditions, it is necessary to start over from the variable selection. Additionally, among different methods for machine learning, the methods better suited than others for developing CBAs have not been addressed yet.

In this study, using a large database of employees' health insurance programs, we developed CBAs with selected machine learning methods for identifying patients with three common chronic conditions: hypertension, diabetes, and dyslipidemia. We then compared their association measures using a hold-out test set.

Methods

Institutional settings

The Japanese government provides a universal health insurance program for all registered inhabitants. Besides, each employer is obliged by law to provide annual health check-up to its employees. Medical and pharmacy claims data combined with annual health check-up results of employees' health insurance programs were obtained in an anonymous format from JMDC Inc. [47]. Further details on the institutional settings have been described previously [33].

Claims data contain enrollee information, including gender, month and year of birth, and their diagnostic code, medical institutions, pharmacies, and medical treatments provided. Diagnostic and medication codes are classified by the 2003 version of the International Statistical Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10) [48] and 2016 version of the World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) [49], respectively. Enrollees' age was defined as their age in March 2018. Annual health check-up results include information on the results of the physical examination and blood test, whether fasting blood samples were collected, and the answer to a health-related questionnaire including questions on medication usage. The study protocol was approved by the research ethics committee of the University of Tokyo (approval number: KE18-44). The ethics committee waived informed consent because this is a retrospective study with the data that were fully anonymized before we accessed them. JMDC Inc. applies strict policies to protect the privacy of enrollees and medical providers, and all private information that could identify enrollees and medical providers were removed beforehand [47].

Study population

The study population for each condition of hypertension, diabetes, and dyslipidemia was defined as beneficiaries (1) who were enrolled in the claims database from April 1, 2016, to March 31, 2018, and whose health check-ups were sequentially conducted for fiscal year (FY) 2016 and FY2017 (N = 1,040,351), (2) with complete data on the self-reported use of blood pressure- and lipid-lowering drugs and hypoglycemic drugs for FY2016 and FY2017 (N = 944,717), (3) who in FY2017 visited a clinic/hospital that mainly specializes in internal medicine (N = 631,731), and (4) with complete data on examination results required for the gold standard of each condition mentioned later for FY2016 and FY2017 (hypertension, N = 631,289; diabetes, N = 152,368; dyslipidemia, N = 614,434) (Fig 1).

In similar studies to date, chart review has often been the source of the gold standard, with the population to calculate association measures constrained to those who visited primary care facilities [15, 16, 19]. To make the present study comparable to the past research, we restricted the study population to those who, at least once in the FY, had visited a clinic/hospital that mainly specializes in internal medicine, which has the function of primary care in Japan.

Gold standard and claims-based algorithm

We constructed a gold standard to diagnose each condition from the health check-up results of FY2016 and FY2017 as previously described (Table 1) [33]. We used FY2017 claims data as the source of the CBA and compared it with the diagnosis derived from the gold standard. The scheme of using one-year claims data corresponding to the latter health check-up year for developing CBAs is the same as that of the previous study [33].

To construct CBAs, we first set up a dataset containing the input variables that can be chosen without subject-matter knowledge on the target conditions, namely, age, gender, and the number of observations of each of ICD-10/WHO-ATC code with a letter followed by two digits (main dataset). We counted the observations of ICD-10/WHO-ATC codes on claims as one occurrence when the information was accrued from the same month. We excluded the ICD-10 codes for suspected cases and counted the ICD-10 codes regardless of whether they were listed as primary diagnoses. We then applied following popular machine learning methods, (1)

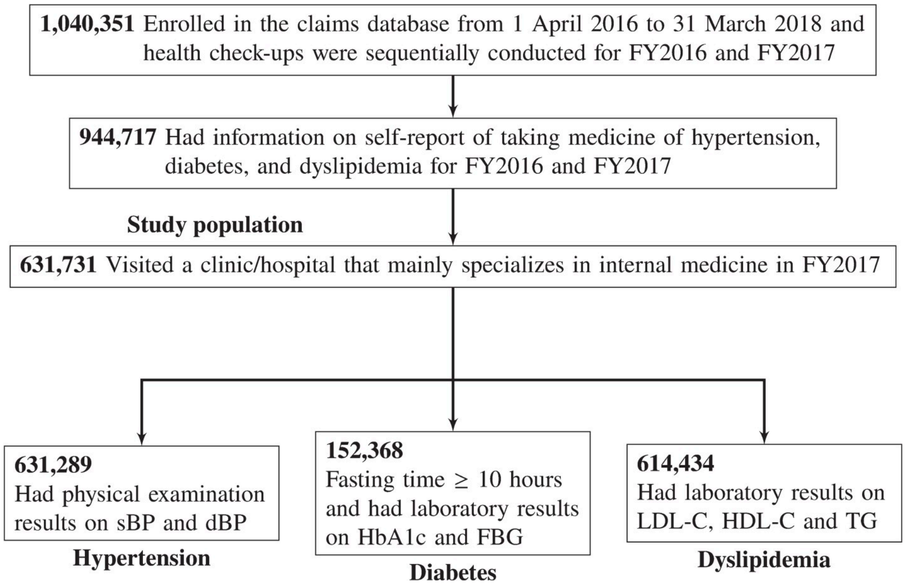
1,040,351 Enrolled in the claims database from 1 April 2016 to 31 March 2018 and health check-ups were sequentially conducted for FY2016 and FY2017
944,717 Had information on self-report of taking medicine of hypertension, diabetes, and dyslipidemia for FY2016 and FY2017
Study population
631,731 Visited a clinic/hospital that mainly specializes in internal medicine in FY2017
631,289 Had physical examination results on sBP and dBP
Hypertension
152,368 Fasting time ≥ 10 hours and had laboratory results on HbA1c and FBG
Diabetes
614,434 Had laboratory results on LDL-C, HDL-C and TG
Dyslipidemia

Fig 1. Flowchart of inclusion and exclusion of study participants. *Abbreviations*: dBP, diastolic blood pressure; FBG, fasting blood glucose; FY, fiscal year; HbA1c, hemoglobin A1c; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; sBP, systolic blood pressure; TG, triglyceride.

https://doi.org/10.1371/journal.pone.0254394.g001

k-nearest neighbor (kNN), (2) support vector machine (SVM), (3) penalized logistic regression, (4) tree-based model, and (5) neural network, to the dataset.

Additionally, as benchmarks, we developed two sets of conventional CBAs. Firstly, we emulated two manually constructed CBAs proposed in the previous study [33]. Patients meeting the following selection rule were classified as "test-positive" for condition X (hypertension, diabetes, or dyslipidemia): (1) the diagnostic code corresponding to condition X is found in the claims at least once (diagnostic code-based CBA); and (2) the medication code corresponding to condition X is found in the claims at least once (medication code-based CBA).

Secondly, we applied a logistic regression model to the main dataset and an alternative dataset where input variables were selected according to each condition. The logistic regression model with the alternative dataset corresponds to a typical procedure among the conventional

Table 1. Gold standards to diagnose hypertension, diabetes, and dyslipidemia.

Diagnose as hypertension if any of the following conditions are satisfied:
1. Systolic blood pressure $\geq$ 140 mmHg and/or diastolic blood pressure $\geq$ 90 mmHg for FY2016 and FY2017
2. Self-report of taking blood pressure-lowering drugs in at least one of FY2016 and FY2017*

Diagnose as diabetes if any of the following conditions are satisfied:
1. HbA1c $\geq$ 6.5% in at least one of the two years and FBG $\geq$ 126 mg/dL in at least one of FY2016 and FY2017
2. FBG $\geq$ 126 mg/dL for FY2016 and FY2017
3. Self-report of taking hypoglycemic drugs in at least one of FY2016 and FY2017*

Diagnose as dyslipidemia if any of the following conditions are satisfied:
1. Low-density lipoprotein cholesterol $\geq$ 140 mg/dL for FY2016 and FY2017
2. High-density lipoprotein cholesterol $\leq$ 40 mg/dL for FY2016 and FY2017
3. Triglyceride $\geq$ 150 mg/dL for FY2016 and FY2017
4. Self-report of taking lipid-lowering drugs in at least one of FY2016 and FY2017*

Abbreviations: HbA1c, hemoglobin A1c; FBG, fasting blood glucose; FY, fiscal year.

*The reliability of the self-report of medication usage was demonstrated to be satisfactorily high when compared with the pharmacy claims-based drug usage [33].

https://doi.org/10.1371/journal.pone.0254394.t001

knowledge-based methods in building CBAs. The selected input variables were age, gender, and the number of observations of each of ICD-10/WHO-ATC codes that corresponds to the target condition. The ICD-10 codes corresponding to hypertension, diabetes, and dyslipidemia were defined as I10-I15, E10-E14, and E78, respectively. The WHO-ATC codes corresponding to hypertension, diabetes, and dyslipidemia were defined as C08 and/or C09, A10, and C10, respectively.

Association measures

We quantified the goodness of CBAs by the following association measures: sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). For the calculation of the association measures, true positive and test positive cases were defined as the enrollees who were assessed as having a disease by the gold standard and those who were identified as having a disease by the CBA, respectively.

Statistical analysis

We randomly divided the dataset into two sets: training (75%), which was used to estimate parameters and tune hyperparameters; test (25%), which was used to assess the association measures of the CBA. The sensitivity, specificity, PPV, and NPV were estimated for the diagnostic code- and medication code-based CBAs, and their 95% confidence intervals (CIs) were calculated using exact binomial confidence limits [50]. We calculated these association measures and 95% CIs using the *epiR* package [51].

We estimated a prediction function that outputs the score of the propensity for having a disease given a set of input variables using the selected methods. The outcome variable in hand is a binary indicator of having a disease that is assessed by the gold standard. For each selected machine learning method, we chose several types of prediction procedures that are commonly applied. The Euclidean distance with raw or standardized (i.e., rescaled to have mean zero and variance one) input variables was adopted as a distance metric for the kNN [52, 53]. A linear basis function with a hinge or squared hinge loss was adopted in the SVM [54]. From the penalized logistic regression, logistic regressions with the L_2-penalty (logistic ridge) [55], L_1-penalty (logistic lasso) [56], and elastic-net penalty (logistic elastic-net) [57] were applied. Two types of tree-based models were applied: random forest [58] and importance sampled learning ensemble (ISLE) [59]. A single hidden layer neural network was applied with a different number of hidden units: 5, 10, and 20 [60].

If the model involved a hyperparameter to be tuned, the training set was used for the tuning. The expected value of the AUC was estimated through tenfold cross-validation with the training set. If the computational burden of tenfold cross-validation was prohibitive, we used a validation set to estimate the expected value of the AUC. A third of the training set was chosen at random to construct the validation set. Which of tenfold cross-validation or a validation set was used for each model is described below. The hyperparameter was then chosen to be the value that maximized the AUC. After the hyperparameter determination, the training set was used again to estimate parameters for the prediction function. When no hyperparameter tuning was required, the training set was used to estimate parameters in the prediction function from the beginning. We described the details of the parameter estimation and hyperparameter tuning for each method in the following.

Logistic regression. The outcome variable was regressed on the input variables to generate a prediction function. The analysis of the logistic regression was implemented by the *mnlogit* package [61].

k-nearest neighbor. The number of the nearest neighbors to be counted, k, was optimized using the validation set. The predicted class probabilities that were computed from (1) the frequency of the class of the k-nearest neighbors (vote) [52] and (2) the inverse distance weighted frequency of the class of the k-nearest neighbors (IDW) [53] composed a prediction function. The analysis of the kNN was implemented by the *fastknn* package [62].

Support vector machine. The cost parameter was optimized using the validation set. Decision values (i.e., the distance of the point from the hyperplane) made up a prediction function. The analysis of the SVM was implemented by the *LiblineaR* package [63].

Penalized logistic regression. The regularization coefficient and elastic-net mixing parameter were determined by cross-validation. The analysis of the penalized logistic regression model was implemented by the *glmnet* package [64].

Tree-based model. The minimum node size was set to 10 for each tree, and 200 trees were bagged in the random forest. The number of variables selected for each split was tuned using the validation set. The probability forest was used to generate a prediction function [65]. The analysis of the random forest was implemented by the *ranger* package [66]. There are five hyperparameters in the importance sampled learning ensemble (ISLE): a hyperparameter for the tree size, subsampling ratio for each tree, learning rate, number of trees to be bagged, and regularization coefficient for the post-processing. We adopted the depth of the tree as the hyperparameter for the tree size and fixed it to be six [60]. As the combination of the subsampling ratio for each tree and learning rate, we selected (1,0.05), (0.5,0.1), and (0.1,0.1). Since the basis function generating process of the ISLE is identical to that of the gradient boosting machine (GBM) if the subsampling ratio is one and that of the stochastic gradient boosting machine (SGBM) if otherwise, we set the learning rate according to Friedman's recommendation for the GBM and SGBM [67, 68]. The remaining two hyperparameters, the number of trees to be bagged and regularization coefficient, were determined by cross-validation. In particular, for a given value of the regularization coefficient, the basis function generating process was stopped if the cross-validation AUC did not improve for three basis function generating rounds. The value with the maximum cross-validation AUC was then chosen as the regularization coefficient for the prediction function. The L_1-penalty was adopted in the post-processing following the recommendation of Friedman and Popescu (2003) [59]. The analysis of the ISLE was implemented by the *xgboost* package [69].

Neural network. All hidden units were fully connected with the nodes in the input and output layers. Weight decay was employed for the regularization of parameters, and the regularization coefficient of it was tuned using the validation set. The analysis of the neural network was implemented by the *nnet* package [70].

Provided an estimated prediction function from the model, an ROC curve was drawn from the scores and the matched observed outcome values as the threshold of considering a patient positive were moved over the range of all possible scores. The AUC was calculated from the resulting ROC curve, and DeLong's method was used to determine the 95% CI for the AUC [71].

In the end, a representative point of sensitivity and specificity on the ROC curve was chosen based on the Youden index [72, 73]. The PPV and NPV were calculated according to the representative point. Moreover, the 95% CIs for the sensitivity, specificity, PPV, and NPV were calculated with 200 bootstrap resampling and the averaging methods as described previously [74]. We drew the ROC curve and calculated the association measures and their 95% CIs using the *pROC* package [75]. All statistical analysis was conducted using R version 3.6.1 [76]. R codes are available at https://github.com/harakonan/research-public/tree/master/cba.

Results

Summary statistics

Table 2 tabulates the summary statistics of 944,717 enrollees' characteristics and health check-up results for each fiscal year. The mean age was 48.0 years (standard deviation ± 10.4 years). More than 80% of people received fasting blood tests. Furthermore, 85% of the enrollees visited any clinics/hospitals during the year, while 67% went to the primary care clinics/hospitals. Among the test cohorts of 157,822, 38,092, and 153,608 enrollees for hypertension, diabetes, and dyslipidemia, 25.4%, 8.4%, and 38.7% of them had a diagnosis of the corresponding condition.

Table 3 displays the cumulative counts and distribution of the proportion of enrollees whose claims contain the ICD-10/WHO-ATC code at least once in the study population. The numbers of the ICD-10 and WHO-ATC codes that appeared in the dataset for the study population were 1333 and 92, respectively. Nearly 90% of the ICD-10 codes that appeared in the dataset were only observed for less than 1% of enrollees, and more than half of the WHO-ATC codes that appeared in the dataset were observed for less than 5% of enrollees.

Table 2. Summary statistics of enrollees' characteristics and health check-up results for each fiscal year of the enrollees with complete data on the self-reported use of blood pressure- and lipid-lowering drugs and hypoglycemic drugs (N = 944,717).

	FY2016			FY2017		
Variables	**Mean**	**SD**	**Missing (%)**	**Mean**	**SD**	**Missing (%)**
Demographics						
Male	–	–	–	0.8	–	–
Age* (year)	–	–	–	48	10.4	–
Visited clinic/hospital						
Any clinic/hospital†	0.85	–	–	0.85	–	–
Primary care clinic/hospital‡	0.67	–	–	0.67	–	–
Health check-up results						
Fasting time ≥ 10 hours$	0.81	–	54.9	0.81	–	56.8
Systolic blood pressure (mmHg)	121.5	15.8	0.1	122.1	15.9	0.0
Diastolic blood pressure (mmHg)	75.5	11.7	0.1	75.9	11.8	0.0
Fasting blood glucose (mg/dL)	96.7	18.5	20.5	97.3	19	21.1
Hemoglobin A1c (%)	5.56	0.64	15.7	5.59	0.64	14.5
Low-density lipoprotein cholesterol (mg/dL)	121.1	30.8	2.4	121.3	30.6	2.8
High-density lipoprotein cholesterol (mg/dL)	60.6	15.9	2.4	60.9	16.1	2.8
Triglyceride (mg/dL)	117.1	94	2.4	118.3	94.5	2.8
Self-report of taking drug¶						
Blood-pressure-lowering drugs	0.12	–	–	0.13	–	–
Hypoglycemic drugs	0.04	–	–	0.04	–	–
Lipid-lowering drugs	0.07	–	–	0.08	–	–

Abbreviations: FY, fiscal year; SD, standard deviation.

Notes: Only mean (or proportion) is stated for a categorical variable. Because the variables "Male" and "Age" do not change with the year, we only tabulated them in column FY2017. There are no missing values in the variables other than the health check-up results by construction.

*Age is defined as the age in March 2018.

†Any clinic/hospital indicates that a person visited any kind of clinic/hospital in the corresponding FY.

‡Primary care clinic/hospital indicates that a person visited a clinic/hospital that mainly provides internal medicine in the corresponding FY.

$Fasting time ≥ 10 hours indicates if more than 10 hours have passed since the last meal when blood samples were collected.

¶Self-report of taking drugs are extracted from the answer to a health-related questionnaire.

https://doi.org/10.1371/journal.pone.0254394.t002

Table 3. Cumulative distribution of the proportion of enrollees whose claims contain the ICD-10/WHO-ATC code at least once in the study population (N = 631,731).

	ICD-10 code		WHO-ATC code	
Proportion	**Count**	**Percentile**	**Count**	**Percentile**
≤ 0.01%	485	36.4 th	5	5.4 th
≤ 0.1%	879	65.9 th	12	13.0 th
≤ 1%	1195	89.6 th	32	34.8 th
≤ 2%	1254	94.1 th	39	42.4 th
≤ 3%	1277	95.8 th	45	48.9 th
≤ 5%	1302	97.7 th	49	53.3 th
≤ 10%	1318	98.9 th	69	75.0 th
≤ 20%	1326	99.5 th	80	87.0 th
≤ 30%	1331	99.8 th	86	93.5 th
≤ 50%	1333	100.0 th	91	98.9 th
≤ 100%	1333	100.0 th	92	100.0 th

Abbreviations: ICD-10, International Classification of Diseases and Related Health Problems, tenth revision; WHO-ATC, World Health Organization-anatomical therapeutic chemical.

Notes: For each two-digit ICD-10/WHO-ATC code, the proportion of enrollees whose claims contain the code at least once was computed for the study population. Cumulative counts and distribution of the computed proportion was tabulated separately for ICD-10 codes and WHO-ATC codes. The count (percentile) column tabulates the number (fraction) of two-digit ICD-10/WHO-ATC codes that the proportion of enrollees whose claims contain the code at least once is below the value in the proportion column.

https://doi.org/10.1371/journal.pone.0254394.t003

Association measures

Table 4 reports the association measures and their 95% CIs for the diagnostic code- and medication code-based CBAs. The sensitivity, specificity, PPV, and NPV closely followed those values computed previously [33]. The diagnostic code-based CBAs had higher sensitivity and NPV but lower specificity and PPV compared to the medication code-based CBAs. For hypertension, all association measures were acceptably high, while, for diabetes, the diagnostic code-

Table 4. Association measures and their 95% confidence intervals for the diagnostic code- and medication code-based claims-based algorithms.

	Sensitivity			Specificity			PPV			NPV		
Method	**%**	**95%CI**		**%**	**95%CI**		**%**	**95%CI**		**%**	**95%CI**	
Hypertension (N = 157,822, Prevalence = 25.4%)												
Diagnostic code	80.7	80.3	81.1	95.2	95.0	95.3	85.0	84.6	85.4	93.6	93.4	93.7
Medication code	75.3	74.9	75.7	97.8	97.7	97.9	92.0	91.7	92.3	92.1	91.9	92.2
Diabetes (N = 38,092, Prevalence = 8.4%)												
Diagnostic code	90.8	89.8	91.8	92.9	92.6	93.2	53.8	52.5	55.2	99.1	99.0	99.2
Medication code	79.3	77.8	80.6	99.5	99.4	99.6	93.5	92.5	94.4	98.1	98.0	98.3
Dyslipidemia (N = 153,608, Prevalence = 38.7%)												
Diagnostic code	49.6	49.2	50.0	90.0	89.8	90.2	75.9	75.5	76.3	73.9	73.6	74.1
Medication code	36.2	35.8	36.6	96.9	96.8	97.0	88.1	87.7	88.5	70.6	70.3	70.8

Abbreviations: CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value.

Patients meeting the following selection rule were classified as “test-positive” for each condition: (1) the diagnostic code corresponding to the condition is found in the claims at least once (diagnostic code-based claims-based algorithm); and (2) the medication code corresponding to the condition is found in the claims at least once (medication code-based claims-based algorithm). We calculated 95% CIs for all estimates of sensitivity, specificity, PPV, and NPV using exact binomial confidence limits.

https://doi.org/10.1371/journal.pone.0254394.t004

based CBA fell short of a satisfactory level of the PPV. For dyslipidemia, the sensitivity of both CBAs was considerably lower than those for hypertension and diabetes.

Table 5 shows the association measures and their 95% CIs for the CBAs derived from the machine learning methods for hypertension (Table 5A), diabetes (Table 5B), and dyslipidemia (Table 5C). ROC curves are shown in S1 File.

The AUC of the logistic regression with subject-matter knowledge about the target condition, i.e., the logistic regression with the alternative dataset, was .923 for hypertension, .957 for diabetes, and .739 for dyslipidemia. The representative sensitivity, specificity, PPV, and NPV of this method were comparable to those of the convex combination of the diagnostic code- and medication code-based CBAs: hypertension, sensitivity 78.0%, specificity 96.1%, PPV 87.0%, and NPV 92.8%; diabetes, 86.9%, 95.6%, 64.3%, and 98.8%; dyslipidemia, 42.6%, 91.8%, 76.6%, and 71.6%. Without subject-matter knowledge about the target condition, i.e., the logistic regression with the main dataset, the AUC for hypertension stayed similar, .921, that for diabetes decreased to .938, and that for dyslipidemia increased to .747.

The logistic lasso, logistic elastic-net, and tree-based methods yielded AUCs that were comparable to or higher than those of the logistic regression with subject-matter knowledge: logistic lasso, .924 for hypertension, .959 for diabetes, and .753 for dyslipidemia; logistic elastic-net, .923, .960, and .747; random forest, .923, .958, and .763; ISLE (the range from three hyperparameter specifications), .930–.931, .964–.966, and .771–.773.

The kNN with raw input variables, SVM, and neural network attained AUCs that were comparable to those of the logistic regression without subject-matter knowledge: kNN with raw input variables (the range from the vote and IDW), .915–.917 for hypertension, .942–.942 for diabetes, and .740–.742 for dyslipidemia; SVM (the range from the hinge and squared hinge loss specifications), .916–.923, .944–.945, and .738–.748; neural network (the range from three different hidden units), .918–.922, .938–.940, and .748–.758.

The kNN with standardized input variables and logistic ridge failed to reach AUCs that were comparable to those of the logistic regression: kNN with standardized input variables (the range from the vote and IDW), .845–.847 for hypertension, .884–.885 for diabetes, and .675–.678 for dyslipidemia; logistic ridge, .892, .928, and .726.

The model which achieved the highest AUC for all three conditions, the ISLE with 1 for the subsampling ratio and 0.05 for the learning rate, yielded the following association measures at the representative coordinate on the ROC curve: hypertension, sensitivity 80.5%, specificity 95.8%, PPV 86.8%, and NPV 93.5%; diabetes, 89.8%, 94.7%, 60.5%, and 99.0%; dyslipidemia, 51.2%, 88.9%, 74.5%, and 74.2%.

Discussion

Using health check-up results as the source of the gold standard, we demonstrated the association measures of the CBAs derived from machine learning methods without a condition-specific variable selection for identifying patients with three common chronic medical conditions, hypertension, diabetes, and dyslipidemia. This is the first study to investigate the benefits of machine learning methods in building CBAs comprehensively.

Among the logistic regression and penalized logistic regression, the logistic lasso and logistic elastic-net achieved the highest AUC, followed by logistic regression and logistic ridge. They are all linear in the parameter model with the same loss function, log-loss, but different penalty functions: zero penalties for logistic regression; an L_2-penalty for logistic ridge; an L_1-penalty for logistic lasso; and an elastic-net penalty for logistic elastic-net.

The methods using the L_1-penalty are better suited to sparse and high-dimensional situations than those using zero penalties or the L_2-penalty because of the selection of the effective

Table 5. Association measures and their 95% confidence intervals for claims-based algorithms derived from machine learning methods.

A. Hypertension (N = 157,822, Prevalence = 25.4%)

	AUC			Sensitivity			Specificity			PPV			NPV		
Method		95%CI		%	95%CI		%	95%CI		%	95%CI		%	95%CI	
Logistic regression															
Main dataset	0.921	0.919	0.923	78.5	78.0	79.1	95.7	95.3	96.1	86.0	85.1	87.1	92.9	92.8	93.1
Alternative dataset	0.923	0.921	0.924	78.0	77.5	78.6	96.1	95.6	96.3	87.0	85.7	87.8	92.8	92.6	92.9
k-nearest neighbor															
Vote	0.917	0.915	0.919	77.2	76.8	77.9	96.1	95.4	96.3	87.2	85.3	87.5	92.6	92.4	92.7
Vote-Standardized	0.847	0.844	0.849	72.7	70.7	73.1	81.8	81.6	83.6	57.6	57.2	59.5	89.8	89.4	90.0
IDW	0.915	0.913	0.917	77.3	76.8	78.0	95.9	95.4	96.2	86.5	85.3	87.3	92.6	92.4	92.7
IDW-Standardized	0.845	0.842	0.847	72.4	70.6	73.5	81.9	80.8	83.6	57.6	56.6	59.5	89.7	89.3	90.0
Support vector machine															
Hinge loss	0.916	0.914	0.918	76.9	76.4	77.7	95.5	94.7	95.8	85.2	83.3	86.2	92.4	92.3	92.6
Squared hinge loss	0.923	0.921	0.924	79.6	79.3	80.2	95.7	95.4	95.9	86.2	85.5	86.8	93.3	93.1	93.4
Penalized logistic regression															
Logistic Ridge	0.892	0.890	0.894	77.1	76.3	77.8	88.0	87.6	88.7	68.7	67.9	69.7	91.9	91.7	92.1
Logistic Lasso	0.924	0.922	0.925	78.7	78.2	79.1	95.6	95.3	96.1	86.0	85.2	87.4	93.0	92.8	93.1
Logistic Elastic-net	0.923	0.921	0.925	78.7	78.2	79.1	95.5	95.2	95.7	85.6	84.8	86.2	92.9	92.8	93.1
Tree-based method															
Random Forest	0.923	0.922	0.925	80.5	79.8	81.1	95.5	95.2	96.1	85.8	85.0	87.5	93.5	93.3	93.7
ISLE-sample 1 learn 0.05	0.931	0.929	0.932	80.5	80.1	81.0	95.8	95.5	96.0	86.8	85.9	87.3	93.5	93.4	93.7
ISLE-sample 0.5 learn 0.1	0.931	0.929	0.932	80.7	80.2	81.1	95.7	95.3	95.9	86.3	85.4	87.0	93.6	93.4	93.7
ISLE-sample 0.1 learn 0.1	0.930	0.928	0.932	80.7	80.2	81.2	95.6	95.2	95.9	86.1	85.2	86.8	93.6	93.4	93.7
Neural network															
Hidden units 5	0.918	0.916	0.919	79.6	78.9	80.2	94.9	94.5	95.4	84.2	83.1	85.4	93.2	93.0	93.4
Hidden units 10	0.922	0.920	0.924	79.4	78.8	80.1	95.6	94.9	96.0	85.9	84.3	87.1	93.2	93.0	93.4
Hidden units 20	0.921	0.919	0.923	79.3	78.8	79.9	95.3	94.7	95.6	85.1	83.6	85.9	93.1	93.0	93.3

B. Diabetes (N = 38,092, Prevalence = 8.4%)

	AUC			Sensitivity			Specificity			PPV			NPV		
Method		95%CI		%	95%CI		%	95%CI		%	95%CI		%	95%CI	
Logistic regression															
Main dataset	0.938	0.932	0.944	85.0	83.4	86.4	95.4	94.1	96.4	62.6	56.8	67.8	98.6	98.5	98.7
Alternative dataset	0.957	0.952	0.961	86.9	85.8	88.4	95.6	94.7	95.9	64.3	59.9	65.9	98.8	98.7	98.9
k-nearest neighbor															
Vote	0.942	0.936	0.948	84.5	82.9	86.2	94.8	93.5	95.8	59.8	54.2	64.2	98.5	98.4	98.7
Vote-Standardized	0.884	0.877	0.891	77.6	75.1	81.9	84.7	80.4	87.3	31.5	27.5	34.7	97.6	97.4	98.0
IDW	0.942	0.936	0.948	84.7	82.9	86.9	95.0	92.9	96.0	60.7	52.5	65.5	98.6	98.4	98.7
IDW-Standardized	0.885	0.878	0.892	78.8	75.8	82.9	83.6	79.8	86.0	30.5	27.0	33.3	97.7	97.5	98.1
Support vector machine															
Hinge loss	0.944	0.938	0.950	86.0	84.3	87.7	95.4	94.0	96.5	62.9	57.0	68.7	98.7	98.5	98.8
Squared hinge loss	0.945	0.939	0.951	86.9	84.9	88.1	95.7	95.0	96.9	64.8	61.5	71.6	98.8	98.6	98.9
Penalized logistic regression															
Logistic Ridge	0.928	0.922	0.933	83.3	80.5	85.1	89.7	88.0	91.8	42.3	39.0	47.2	98.3	98.1	98.5
Logistic Lasso	0.959	0.955	0.964	88.7	87.4	89.8	94.8	94.3	95.4	60.8	58.5	63.5	98.9	98.8	99.0
Logistic Elastic-net	0.960	0.956	0.964	88.9	87.8	89.9	94.5	93.9	95.0	59.7	57.0	61.5	98.9	98.8	99.0
Tree-based method															
Random Forest	0.958	0.953	0.962	88.5	86.5	90.0	94.9	93.5	96.6	61.3	55.8	70.0	98.9	98.7	99.0
ISLE-sample 1 learn 0.05	0.966	0.962	0.970	89.8	88.5	90.9	94.7	93.8	95.1	60.5	56.9	62.4	99.0	98.9	99.1
ISLE-sample 0.5 learn 0.1	0.965	0.961	0.969	89.8	88.7	91.2	94.5	93.5	95.0	59.9	56.2	61.9	99.0	98.9	99.2
ISLE-sample 0.1 learn 0.1	0.964	0.960	0.968	90.3	88.3	91.4	93.7	92.9	95.4	56.5	53.9	64.1	99.1	98.9	99.2
Neural network															
Hidden units 5	0.938	0.932	0.944	83.3	81.2	85.0	94.5	93.1	96.1	57.9	52.6	65.8	98.4	98.2	98.6
Hidden units 10	0.940	0.935	0.946	83.7	81.9	86.0	95.4	93.3	96.4	62.5	53.6	67.7	98.5	98.3	98.6

(*Continued*)

Table 5. (Continued)

Hidden units 20	0.938	0.932	0.944	83.9	82.3	85.8	95.8	94.2	97.0	64.7	57.1	71.5	98.5	98.3	98.6

C. Dyslipidemia (N = 153,608, Prevalence = 38.7%)

	AUC			Sensitivity			Specificity			PPV			NPV		
Method		95%CI		%	95%CI		%	95%CI		%	95%CI		%	95%CI	
Logistic regression															
Main dataset	0.747	0.744	0.749	48.8	43.8	51.1	86.1	83.9	91.0	69.2	66.7	75.5	72.6	71.9	73.1
Alternative dataset	0.739	0.736	0.742	42.6	42.1	43.4	91.8	91.1	92.1	76.6	75.3	77.2	71.6	71.5	71.8
k-nearest neighbor															
Vote	0.742	0.739	0.745	46.6	45.1	49.5	89.3	86.4	90.4	73.2	69.5	75.0	72.6	72.3	73.0
Vote-Standardized	0.678	0.676	0.681	54.6	47.6	55.4	69.5	69.0	76.4	53.2	52.7	56.1	70.7	69.7	71.0
IDW	0.740	0.737	0.742	48.2	45.3	50.0	87.3	85.7	90.4	70.9	69.0	74.7	72.7	72.3	73.0
IDW-Standardized	0.675	0.673	0.678	53.8	49.3	54.9	70.3	69.3	74.6	53.4	52.9	55.3	70.6	69.9	70.9
Support vector machine															
Hinge loss	0.738	0.735	0.740	49.6	49.0	50.0	90.0	89.8	90.2	75.8	75.4	76.2	73.8	73.6	74.0
Squared hinge loss	0.748	0.746	0.751	50.0	45.7	50.6	85.2	84.9	89.3	68.2	67.7	73.0	72.9	72.2	73.1
Penalized logistic regression															
Logistic Ridge	0.726	0.723	0.728	54.8	51.2	58.2	76.6	73.2	80.2	59.8	57.9	62.0	72.8	72.2	73.5
Logistic Lasso	0.753	0.751	0.756	49.3	48.9	49.7	90.3	90.0	90.5	76.3	75.8	76.7	73.8	73.6	74.0
Logistic Elastic-net	0.747	0.744	0.749	45.5	44.5	47.4	89.7	87.7	90.4	73.5	70.8	74.7	72.2	72.0	72.6
Tree-based method															
Random Forest	0.763	0.761	0.766	49.8	48.3	50.3	90.0	89.7	91.4	76.0	75.4	78.0	73.9	73.6	74.1
ISLE-sample 1 learn 0.05	0.773	0.771	0.775	51.2	49.9	52.0	88.9	88.0	90.1	74.5	73.2	76.2	74.2	74.0	74.4
ISLE-sample 0.5 learn 0.1	0.773	0.770	0.775	51.4	49.5	52.5	88.8	87.5	90.6	74.3	72.7	76.9	74.2	73.9	74.5
ISLE-sample 0.1 learn 0.1	0.771	0.768	0.773	50.0	49.0	51.5	90.0	88.4	90.9	75.9	73.6	77.4	74.0	73.8	74.3
Neural network															
Hidden units 5	0.748	0.745	0.750	48.3	46.6	50.8	87.7	85.3	89.4	71.3	68.5	73.5	72.9	72.5	73.3
Hidden units 10	0.751	0.749	0.754	47.7	46.7	50.0	89.3	86.8	90.2	73.9	70.6	75.2	73.0	72.8	73.3
Hidden units 20	0.758	0.755	0.760	49.4	48.3	51.8	88.5	86.0	89.5	73.0	70.1	74.5	73.5	73.2	73.8

Abbreviations: AUC, area under the receiver operating characteristic curve; CI, confidence interval; IDW, inverse distance weighting; ISLE, importance sampled learning ensemble; NPV, negative predictive value; PPV, positive predictive value.

Notes: Age, gender, and all International Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)/World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) codes with a letter followed by two digits were used as input variables for all models but the logistic regression using the alternative dataset. The main logistic regression fitted a logistic regression model to the dataset that was appropriately trimmed. The Euclidean distance with raw or standardized (i.e., rescaled to have mean zero and variance one) input variables was adopted as a distance metric for the k-nearest neighbor (kNN). The number of the nearest neighbors to be counted, k, was optimized using the validation set. The predicted class probabilities that were computed from (1) the frequency of the class of the k-nearest neighbors (vote) and (2) the inverse distance weighted frequency of the class of the k-nearest neighbors (IDW) composed a prediction function. A linear basis function with a hinge or squared hinge loss was adopted in the support vector machine (SVM). The cost parameter was optimized using the validation set. Decision values (i.e., the distance of the point from the hyperplane) made up a prediction function. From the penalized logistic regression, logistic regression with the L_2-penalty (logistic ridge), L_1-penalty (logistic lasso), and elastic-net penalty (logistic elastic-net) were applied. The regularization coefficient and elastic-net mixing parameter were determined by cross-validation. Two types of tree-based models were applied: random forest and importance sampled learning ensemble (ISLE). The minimum node size was set to 10 for each tree, and 200 trees were bagged in the random forest. The number of variables selected for each split was tuned using the validation set. We fixed the depth to be six for the ISLEs. As the combination of the subsampling ratio for each tree and the learning rate, we selected (1,0.05), (0.5,0.1), and (0.1,0.1). The number of trees and regularization coefficient were determined by cross-validation. The L_1-penalty was adopted in the post-processing. A single hidden layer neural network was applied with a different number of hidden units: 5, 10, and 20. All hidden units were fully connected with the nodes in the input and output layers. Weight decay was employed for the regularization of parameters, and the regularization coefficient of it was tuned using the validation set. Delong's method was used to determine the 95% CI for the AUC. A representative point of sensitivity and specificity on the ROC curve is chosen based on the Youden index. PPV and NPV were calculated according to the representative point, and the 95% CIs for the resulting sensitivity, specificity, PPV, and NPV were calculated with 200 bootstrap resampling and the averaging methods.

https://doi.org/10.1371/journal.pone.0254394.t005

input variables. These results are backed by theoretical results that support the superiority of the estimation methods that use the L_1-penalty in sparse and high-dimensional settings [77–79]. Despite the fact that the prediction performance of the lasso is expected to be improved by the elastic-net if there is a group of variables among which the pairwise correlations are very high [46] and usually the diagnostic and medication codes corresponding to the target disease are highly correlated, we could not boost the AUC by the elastic-net compared with the lasso.

The tree-based model and neural network automatically select the input variables that are crucial to the discrimination and flexibly incorporate nonlinearity and interactions of them. The tree-based model largely attained superior AUC to any models and was at least as good as the benchmark cases. Among the tree-based models, the ISLE performed better than the random forest. Past Monte Carlo simulation studies have shown the superior performance of the ISLE to the random forest that uses the lasso post-processing in the aggregation process and the superior performance of the latter to the usual random forest [59, 60]. Therefore, two components of the ISLE are contributing to its superior performance to that of the random forest: learning term in the basis function generating process and lasso post-processing. The difference of the hyperparameter within the ISLE was not so much affecting the results.

In contrast to the tree-based model, the AUC of the neural network was not that high but comparable to that of the logistic regression. The performance of the neural network was much lower in the preliminary investigation that used a smaller sample size. The number of parameters in the neural network is nearly 7500, 15000, and 30000 for 5, 10, and 20 hidden units, respectively. Although the use of weight decay should alleviate the overfitting of the parameters to some extent, the sample size may still be insufficient for the neural network to demonstrate its true predictive power. As using multiple hidden layers with constraints such as local connectivity and weight sharing on the network, which allow for more complex connectivity but fewer parameters, improved the performance of the neural network dramatically in the field of image recognition [80, 81], it may also improve the performance of the neural network in the current subject. Increasing the sample size of data and devising more complex connectivity that suits the situation are fruitful directions for future research.

The AUC of the kNN with raw input variables was as good as that of the logistic regression, but that of the kNN with standardized input variables was lower. As is implicated by the difference of the AUC of the kNN with raw and standardized input variables, designing the distance metric in the kNN is difficult. If the input variables are standardized, the model is coerced to attach less importance on the input variables with high standard deviation, such as age and gender, than otherwise. Although the kNN had established an era in image recognition by the invention of the tangent distance [82], there is no such versatile distance measure yet in the field of CBA or studies using administrative data. It may be possible to improve the performance of the kNN by applying an unsupervised learning method that extracts essential components of the input variables, for example, principal component analysis [83], before measuring the distance. Although we do not probe further in this study, this is one direction of future research.

The AUC of the SVM was higher than that of the logistic ridge. They are linear in the parameter model with the same L_2-penalty but different loss functions. The logistic ridge uses the log-loss, while the SVM uses the (squared) hinge loss. The hinge losses give zero penalties to points correctly classified and outside the margin. On the other hand, the log-loss gives continuously decreasing penalties as the correctly classified points get farther from the boundary of the margin. This feature of the hinge losses makes the SVM more robust to outliers than the other methods that are using the log-loss. Since most of the enrollees were far from the margin or outliers (i.e., most of them could be easily labeled as disease or non-disease by the CBA), the

SVM is achieving a higher performance by better-discriminating enrollees with and without the target disease near the boundary than the other methods.

By comparing the results from the two datasets prepared for the logistic regression, we can see that the AUC declined for diabetes without a condition-specific variable selection. The inconsistency of the trend of the AUC among the target conditions demonstrates the trade-off between the accuracy and variance of the prediction function. When the number of the input variables of the prediction model becomes large relative to the sample size, there is a potential accuracy gain from the use of rich information and a possibility of a variance increase due to the variance inflation of the parameter estimates. In diabetes, the main dataset did not provide enough accuracy gain to offset the variance inflation, as the sample size was relatively small and the factors of being diagnosed as the target condition are successfully captured in the condition-specific variable selection (i.e., a high AUC is achieved by the alternative dataset). Conversely, in dyslipidemia, since the factors of being diagnosed as the condition seem to be not sufficiently covered by the condition-specific variable selection, the effect of the accuracy gain outweighs that of the variance increase.

There are potentially various ways of refining the AUC obtained in this study, drawing on the context of machine learning. Although the objective of this study is not to seek high AUC or prediction accuracy but to outline the prospect of the development of an efficient CBA construction procedure, we briefly introduce the concepts that are expected to become significant in the future accuracy pursuit of CBAs. The first one is more complicated and sophisticated learning models flourished in the field of machine learning, such as deep learning models [84]. The second is the pre-processing techniques that transform datasets *ex-ante* to utilize the power of learning machines more efficiently. There are mainly two approaches for pre-processing: methods that deal with imbalanced datasets [85] and those that perform feature selection [86]. The last one is error analysis in the performance analysis and debugging step of model building [87]. How one can successfully use these methods in CBA or, more broadly, claims data situation should be a worthwhile subject to be pursued.

We note that, admittedly, the most demanding and time-consuming task when conducting CBA research will usually be the construction of gold standards. For instance, most previous studies reviewed medical charts to construct the gold standard [13, 15, 16, 18, 19, 21–23, 25, 26, 28, 36, 39, 41]. Nevertheless, we still believe that our proposed method may also lower the bar of CBA research and useful for the following three reasons. Firstly, the performance measures calculated using appropriate machine learning methods can be potentially useful as a reference point, even when creating CBAs manually or exploring new CBA construction procedures.

Secondly, in some cases, it may be possible to sidestep the burden of chart reviewing by using regularly collected data like annual health screening results, which are used in this study. Electronic medical records and disease registries are possible candidates along this line. An increasing number of phenotype algorithms [88–90] may well function as gold standards for CBA research when electronic medical records are available. Cancer registries can be used to conduct comprehensive CBA research for various cancers. In fact, some CBA research is using health screening results [27, 33], blood test results from electronic medical records [22, 31], and disease registries [9, 14, 17, 24, 29, 40]. If this is the case, researchers can construct the gold standards from the regularly collected data without a serious burden and may be able to apply our proposed method to construct CBAs for a broad set of target conditions once an initial set of input variables are selected.

Finally, as we highlighted in the Introduction, there are demands to build CBAs for a wide variety of diseases efficiently. For instance, we need to renew CBAs when the coding scheme changes [42, 43], and a number of countries are still suffering from a lack of CBAs [44]. The lack of confirmed CBAs degrades the research quality, and the research will likely fail to attract high impact factor journals' attention [45].

There is a two-dimensional generalizability issue on the value of association measures computed here: the study population only covers regular employees; the research only dealt with three conditions, hypertension, diabetes, and dyslipidemia. Additionally, input variables selected without using subject-matter knowledge on the target conditions in this study may be inadequate for other situations and conditions. Additional enrollee characteristics, ICD-10/WHO-ATC codes with three or more digits, and procedure codes may need to be included to attain satisfactory CBAs. The information on primary diagnoses and suspected cases may also be helpful. However, considering that comorbidities were our focus, we do not expect that incorporating those types of information would appreciably affect the methods' accuracy in this study. Lastly, the learning method that suits may depend on the target condition. We hope that similar studies will be conducted on situations other than those that were investigated in the present research to gain a deeper understanding regarding the development of efficient CBA research.

In sum, the penalized logistic regressions other than ridge and tree-based models, which are the leading machine learning methods, achieved AUCs comparable to the logistic regression with a knowledge-based condition-specific variable selection. Besides, the AUC level was satisfactory for hypertension and diabetes. Appropriate machine learning methods can substitute our knowledge of target conditions to construct CBAs efficiently.

Supporting information

S1 File. Receiver operating characteristic curve for claims-based algorithms derived from machine learning methods (A, Hypertension; B, Diabetes; C, Dyslipidemia). *Abbreviations*: AUC, area under the receiver operating characteristic curve; IDW, inverse distance weighting; ISLE, importance sampled learning ensemble; kNN, k-nearest neighbor; Std., standardized; SVM, support vector machine; RF, random forest. Notes: Age, gender, and all International Classification of Diseases and Related Health Problems, Tenth Revision (ICD-10)/World Health Organization-Anatomical Therapeutic Chemical (WHO-ATC) codes with a letter followed by two digits were used as input variables for all models but the logistic regression using the alternative dataset. The main logistic regression fitted a logistic regression model to the dataset that was appropriately trimmed. The Euclidean distance with raw or standardized (i.e., rescaled to have mean zero and variance one) input variables was adopted as a distance metric for the k-nearest neighbor (kNN). The number of the nearest neighbors to be counted, k, was optimized using the validation set. The predicted class probabilities that were computed from (1) the frequency of the class of the k-nearest neighbors (vote) and (2) the inverse distance weighted frequency of the class of the k-nearest neighbors (IDW) composed a prediction function. A linear basis function with a hinge or squared hinge loss was adopted in the support vector machine (SVM). The cost parameter was optimized using the validation set. Decision values (i.e., the distance of the point from the hyperplane) made up a prediction function. From the penalized regression, logistic regression with the L_2-penalty (logistic ridge), L_1-penalty (logistic lasso), and elastic-net penalty (logistic elastic-net) were applied. The regularization coefficient and elastic-net mixing parameter were determined by cross-validation. Two types of tree-based models were applied: random forest and importance sampled learning ensemble (ISLE). The minimum node size was set to 10 for each tree, and 200 trees were bagged in the random forest. The number of variables selected for each split was tuned using the validation set. We fixed the depth to be six for the ISLEs. As the combination of the subsampling ratio for each tree and the learning rate, we selected (1,0.05), (0.5,0.1), and (0.1,0.1).

The number of trees and regularization coefficient were determined by cross-validation. The L_1-penalty was adopted in the post-processing. A single hidden layer neural network was applied with a different number of hidden units: 5, 10, and 20. All hidden units were fully connected with the nodes in the input and output layers. Weight decay was employed for the regularization of parameters, and the regularization coefficient of it was tuned using the validation set.
(DOCX)

Acknowledgments

We appreciate Masahiro Nakamura for his management activities regarding the computing environment.

Author Contributions

Conceptualization: Konan Hara.

Formal analysis: Konan Hara, Ryo Ikesu.

Funding acquisition: Ung-il Chung, Akiko Kishi Svensson.

Methodology: Konan Hara, Yuki Ito.

Project administration: Ung-il Chung, Akiko Kishi Svensson.

Resources: Akiko Kishi Svensson.

Software: Konan Hara.

Supervision: Yasuki Kobayashi, Jun Tomio.

Visualization: Konan Hara.

Writing – original draft: Konan Hara.

Writing – review & editing: Yasuki Kobayashi, Jun Tomio, Yuki Ito, Thomas Svensson, Ryo Ikesu, Ung-il Chung, Akiko Kishi Svensson.

References

1. Iizuka T. Physician agency and adoption of generic pharmaceuticals. Am Econ Rev. 2012; 102: 2826–2858. https://doi.org/10.1257/aer.102.6.2826 PMID: 29522299
2. Einav L, Finkelstein A, Schrimpf P. The Response of Drug Expenditure to Contract Design in Medicare Part D. Q J Econ. 2015; 130: 841–899. https://doi.org/10.1093/qje/qjv005 PMID: 26769984
3. Schermerhorn ML, Buck DB, O'Malley AJ, Curran T, McCallum JC, Darling J, et al. Long-Term Outcomes of Abdominal Aortic Aneurysm in the Medicare Population. N Engl J Med. 2015; 373: 328–338. https://doi.org/10.1056/NEJMoa1405778 PMID: 26200979
4. Abaluck J, Gruber J. Evolving choice inconsistencies in choice of prescription drug insurance. Am Econ Rev. 2016; 106: 2145–2184. https://doi.org/10.1257/aer.20130778 PMID: 29104294
5. McWilliams JM, Hatfield LA, Chernew ME, Landon BE, Schwartz AL. Early Performance of Accountable Care Organizations in Medicare. N Engl J Med. 2016; 374: 2357–2366. https://doi.org/10.1056/NEJMsa1600142 PMID: 27075832
6. Nuti S V., Qin L, Rumsfeld JS, Ross JS, Masoudi FA, Normand S-LT, et al. Association of Admission to Veterans Affairs Hospitals vs Non-Veterans Affairs Hospitals With Mortality and Readmission Rates Among Older Men Hospitalized With Acute Myocardial Infarction, Heart Failure, or Pneumonia. JAMA. 2016; 315: 582–92. https://doi.org/10.1001/jama.2016.0278 PMID: 26864412
7. Layton JB, Kim Y, Alexander GC, Emery SL. Association Between Direct-to-Consumer Advertising and Testosterone Testing and Initiation in the United States, 2009–2013. JAMA. 2017; 317: 1159–1166. https://doi.org/10.1001/jama.2016.21041 PMID: 28324090

8. Virnig BA, McBean M. Administrative Data for Public Health Surveillance and Planning. Annu Rev Public Health. 2001; 22: 213–230. https://doi.org/10.1146/annurev.publhealth.22.1.213 PMID: 11274519

9. Taylor DH, Fillenbaum GG, Ezell ME. The accuracy of medicare claims data in identifying Alzheimer's disease. J Clin Epidemiol. 2002; 55: 929–937. https://doi.org/10.1016/s0895-4356(02)00452-3 PMID: 12393082

10. Rector TS, Wickstrom SL, Shah M, Greeenlee NT, Rheault P, Rogowski J, et al. Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions. Health Serv Res. 2004; 39: 1839–1857. https://doi.org/10.1111/j.1475-6773.2004.00321.x PMID: 15533190

11. Kern EFO, Maney M, Miller DR, Tseng C-L, Tiwari A, Rajan M, et al. Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes. Health Serv Res. 2006; 41: 564–580. https://doi.org/10.1111/j.1475-6773.2005.00482.x PMID: 16584465

12. Klabunde CN, Harlan LC, Warren JL. Data sources for measuring comorbidity: a comparison of hospital records and medicare claims for cancer patients. Med Care. 2006; 44: 921–8. https://doi.org/10.1097/01.mlr.0000223480.52713.b9 PMID: 17001263

13. Losina E, Barrett J, Baron JA, Katz JN. Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients. J Clin Epidemiol. 2003; 56: 515–519. https://doi.org/10.1016/s0895-4356(03)00056-8 PMID: 12873645

14. Nattinger AB, Laud PW, Bajorunaite R, Sparapani RA, Freeman JL. An Algorithm for the Use of Medicare Claims Data to Identify Women with Incident Breast Cancer. Health Serv Res. 2004; 39: 1733–1750. https://doi.org/10.1111/j.1475-6773.2004.00315.x PMID: 15533184

15. Wilchesky M, Tamblyn RM, Huang A. Validation of diagnostic codes within medical services claims. J Clin Epidemiol. 2004; 57: 131–141. https://doi.org/10.1016/S0895-4356(03)00246-4 PMID: 15125622

16. Bullano MF, Kamat S, Willey VJ, Barlas S, Watson DJ, Brenneman SK. Agreement Between Administrative Claims and the Medical Record in Identifying Patients With a Diagnosis of Hypertension. Med Care. 2006; 44: 486–490. https://doi.org/10.1097/01.mlr.0000207482.02503.55 PMID: 16641668

17. Gold HT, Do HT. Evaluation of three algorithms to identify incident breast cancer in medicare claims data. Health Serv Res. 2007; 42: 2056–2069. https://doi.org/10.1111/j.1475-6773.2007.00705.x PMID: 17850533

18. Nordstrom BL, Norman HS, Dube TJ, Wilcox MA, Walker AM. Identification of abacavir hypersensitivity reaction in health care claims data. Pharmacoepidemiol Drug Saf. 2007; 16: 289–296. https://doi.org/10.1002/pds.1337 PMID: 17245797

19. Quan H, Khan N, Hemmelgarn BR, Tu K, Chen G, Campbell N, et al. Validation of a case definition to define hypertension using administrative data. Hypertension. 2009; 54: 1423–1428. https://doi.org/10.1161/HYPERTENSIONAHA.109.139279 PMID: 19858407

20. Taylor DH, Østbye T, Langa KM, Weir D, Plassman BL. The accuracy of medicare claims as an epidemiological tool: The case of dementia revisited. J Alzheimer's Dis. 2009; 17: 807–815. https://doi.org/10.3233/JAD-2009-1099 PMID: 19542620

21. Cheng C-L, Kao Y-HY, Lin S-J, Lee C-H, Lai ML. Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan. Pharmacoepidemiol Drug Saf. 2011; 20: 236–42. https://doi.org/10.1002/pds.2087 PMID: 21351304

22. Gorina Y, Kramarow EA. Identifying chronic conditions in medicare claims data: Evaluating the chronic condition data warehouse algorithm. Health Serv Res. 2011; 46: 1610–1627. https://doi.org/10.1111/j.1475-6773.2011.01277.x PMID: 21649659

23. Quam L, Ellis LBM, Venus P, Clouse J, Taylor CG, Leatherman S. Using claims data for epidemiologic research. The concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population. Med Care. 1993; 31: 498–507. PMID: 8501997

24. Kawasumi Y, Abrahamowicz M, Ernst P, Tamblyn R. Development and validation of a predictive algorithm to identify adult asthmatics from medical services and pharmacy claims databases. Health Serv Res. 2011; 46: 939–963. https://doi.org/10.1111/j.1475-6773.2010.01235.x PMID: 21275988

25. Scholes D, Yu O, Raebel MA, Trabert B, Holt VL. Improving automated case finding for ectopic pregnancy using a classification algorithm. Hum Reprod. 2011; 26: 3163–3168. https://doi.org/10.1093/humrep/der299 PMID: 21911435

26. Tu K, Manuel D, Lam K, Kavanagh D, Mitiku TF, Guo H. Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions. J Clin Epidemiol. 2011; 64: 431–435. https://doi.org/10.1016/j.jclinepi.2010.04.007 PMID: 20638237

27. Tessier-Sherman B, Galusha D, Taiwo OA, Cantley L, Slade MD, Kirsche SR, et al. Further validation that claims data are a useful tool for epidemiologic research on hypertension. BMC Public Health. 2013; 13: 51. https://doi.org/10.1186/1471-2458-13-51 PMID: 23331960

28. Cheng C-L, Lee C-H, Chen P-S, Li Y-H, Lin S-J, Yang Y-HK. Validation of Acute Myocardial Infarction Cases in the National Health Insurance Research Database in Taiwan. 2014; 24: 500–507. https://doi.org/10.2188/jea.je20140076 PMID: 25174915

29. Chan A-W, Fung K, Tran JM, Kitchen J, Austin PC, Weinstock MA, et al. Application of Recursive Partitioning to Derive and Validate a Claims-Based Algorithm for Identifying Keratinocyte Carcinoma (Nonmelanoma Skin Cancer). JAMA Dermatology. 2016; 152: 1122. https://doi.org/10.1001/jamadermatol.2016.2609 PMID: 27533718

30. van Walraven C, Colman I. Migraineurs were reliably identified using administrative data. J Clin Epidemiol. 2016; 71: 68–75. https://doi.org/10.1016/j.jclinepi.2015.09.007 PMID: 26404461

31. Yamana H, Horiguchi H, Fushimi K, Yasunaga H. Comparison of Procedure-Based and Diagnosis-Based Identifications of Severe Sepsis and Disseminated Intravascular Coagulation in Administrative Data. J Epidemiol. 2016; 26: 1–8. https://doi.org/10.2188/jea.JE20150286 PMID: 27064132

32. Yamana H, Moriwaki M, Horiguchi H, Kodan M, Fushimi K, Yasunaga H. Validity of diagnoses, procedures, and laboratory data in Japanese administrative data. J Epidemiol. 2017; 1–7. https://doi.org/10.1016/j.je.2016.09.009 PMID: 28142051

33. Hara K, Tomio J, Svensson T, Ohkuma R, Svensson AK, Yamazaki T. Association measures of claims-based algorithms for common chronic conditions were assessed using regularly collected data in Japan. J Clin Epidemiol. 2018; 99: 84–95. https://doi.org/10.1016/j.jclinepi.2018.03.004 PMID: 29548842

34. Hebert PL, Geiss LS, Tierney EF, Engelgau MM, Yawn BP, McBean AM. Identifying Persons with Diabetes Using Medicare Claims Data. Am J Med Qual. 1999; 14: 270–277. https://doi.org/10.1177/106286069901400607 PMID: 10624032

35. Østbye T, Taylor DH, Clipp EC, Scoyoc L Van, Plassman BL. Identification of dementia: Agreement among national survey data, medicare claims, and death certificates. Health Serv Res. 2008; 43: 313–326. https://doi.org/10.1111/j.1475-6773.2007.00748.x PMID: 18211532

36. Katz JN, Barrett J, Liang MH, Bacon AM, Kaplan H, Kieval RI, et al. Sensitivity and positive predictive value of medicare part B physician claims for rheumatologic diagnoses and procedures. Arthritis Rheum. 1997; 40: 1594–1600. https://doi.org/10.1002/art.1780400908 PMID: 9324013

37. Muhajarine N, Mustard C, Roos LL, Young TK, Gelskey DE. Comparison of survey and physician claims data for detecting hypertension. J Clin Epidemiol. 1997; 50: 711–718. https://doi.org/10.1016/s0895-4356(97)00019-x PMID: 9250269

38. Robinson JR, Young TK, Roos LL, Gelskey DE. Estimating the burden of disease. Comparing administrative data and self-reports. Med Care. 1997; 35: 932–947. https://doi.org/10.1097/00005650-199709000-00006 PMID: 9298082

39. Sands K, Vineyard G, Livingston J, Christiansen C, Platt R. Efficient Identification of Postdischarge Surgical Site Infections: Use of Automated Pharmacy Dispensing Information, Administrative Data, and Medical Record Information. J Infect Dis. 1999; 179: 434–441. https://doi.org/10.1086/314586 PMID: 9878028

40. Freeman JL, Zhang D, Freeman DH, Goodwin JS. An approach to identifying incident breast cancer cases using Medicare claims data. J Clin Epidemiol. 2000; 53: 605–614. https://doi.org/10.1016/s0895-4356(99)00173-0 PMID: 10880779

41. Andrade SE, Gurwitz JH, Chan KA, Donahue JG, Beck A, Boles M, et al. Validation of diagnoses of peptic ulcers and bleeding from administrative databases: A multi-health maintenance organization study. J Clin Epidemiol. 2002; 55: 310–313. https://doi.org/10.1016/s0895-4356(01)00480-2 PMID: 11864803

42. Link J, Glazer C, Torres F, Chin K. International Classification of Diseases Coding Changes Lead to Profound Declines in Reported Idiopathic Pulmonary Arterial Hypertension Mortality and Hospitalizations. Chest. 2011; 139: 497–504. https://doi.org/10.1378/chest.10-0837 PMID: 20724737

43. Khera R, Dorsey KB, Krumholz HM. Transition to the ICD-10 in the United States. JAMA. 2018; 320: 133. https://doi.org/10.1001/jama.2018.6823 PMID: 29868861

44. Koram N, Delgado M, Stark JH, Setoguchi S, Luise C. Validation studies of claims data in the Asia-Pacific region: A comprehensive review. Pharmacoepidemiol Drug Saf. 2019; 28: 156–170. https://doi.org/10.1002/pds.4616 PMID: 30022560

45. Van Walraven C, Bennett C, Forster AJ. Administrative database research infrequently used validated diagnostic or procedural codes. J Clin Epidemiol. 2011; 64: 1054–1059. https://doi.org/10.1016/j.jclinepi.2011.01.001 PMID: 21474278

46. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodol. 2005; 67: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

47. Kimura S, Sato T, Ikeda S, Noda M, Nakayama T. Development of a Database of Health Insurance Claims: Standardization of Disease Classifications and Anonymous Record Linkage. J Epidemiol. 2010; 20: 413–419. https://doi.org/10.2188/jea.je20090066 PMID: 20699602

48. WHO. WHO—International Classification of Diseases. [cited 10 Oct 2018]. Available: http://www.who.int/classifications/icd/en/

49. WHO. WHOCC—ATC/DDD Index. [cited 10 Oct 2018]. Available: https://www.whocc.no/atc_ddd_index/

50. Collet D. Modelling Binary Data. Second. Boca Raton, Florida: Chapman & Hall/CRC; 1999.

51. Stevenson M, Nunes T, Heuer C, Marshall J, Sanchez J, Thornton R, et al. epiR: Tools for the Analysis of Epidemiological Data. 2018.

52. Fix E, Hodges JL. Discriminatory Analysis—Nonparametric Discrimination: Consistency Properties. 1951.

53. Shepard D. A two-dimensional interpolation function for irregularly-spaced data. Proceedings of the 23rd ACM National Conference. New York, USA: ACM; 1968. pp. 517–524. https://doi.org/10.1145/800186.810616

54. Cristianini N, Shawe-Taylor J. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. New York: Cambridge University Press; 2000.

55. Zhu J, Hastie T. Classification of gene microarrays by penalized logistic regression. Biostatistics. 2004; 5: 427–443. https://doi.org/10.1093/biostatistics/5.3.427 PMID: 15208204

56. Shevade SK, Keerthi SS. A simple and efficient algorithm for gene selection using sparse logistic regression. Bioinformatics. 2003; 19: 2246–2253. https://doi.org/10.1093/bioinformatics/btg308 PMID: 14630653

57. Waldron L, Pintilie M, Tsao M, Shepherd FA, Huttenhower C, Jurisica I. Optimized application of penalized regression methods to diverse genomic data. Bioinformatics. 2011; 27: 3399–3406. https://doi.org/10.1093/bioinformatics/btr591 PMID: 22156367

58. Breiman L. Random Forests. Mach Learn. 2001; 45: 5–32. https://doi.org/10.1023/A:1010933404324

59. Friedman JH, Popescu BE. Importance Sampled Learning Ensembles. 2003.

60. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: Springer; 2009. https://doi.org/10.1007/978-0-387-84858-7

61. Hasan A, Wang Z, Mahani AS. Fast Estimation of Multinomial Logit Models: R Package mnlogit. J Stat Softw. 2016; 75. https://doi.org/10.18637/jss.v075.i03

62. Pinto D. fastknn: Build Fast k-Nearest Neighbor Classifiers. 2018.

63. Helleputte T. LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library. 2017.

64. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2010; 33: 1–22. PMID: 20808728

65. Malley JD, Kruppa J, Dasgupta A, Malley KG, Ziegler A. Probability Machines. Methods Inf Med. 2012; 51: 74–81. https://doi.org/10.3414/ME00-01-0052 PMID: 21915433

66. Wright MN, Ziegler A. ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. J Stat Softw. 2017; 77. https://doi.org/10.18637/jss.v077.c01 PMID: 28649186

67. Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat. 2001; 29: 1189–1232. https://doi.org/10.1214/aos/1013203451

68. Friedman JH. Stochastic gradient boosting. Comput Stat Data Anal. 2002; 38: 367–378. https://doi.org/10.1016/S0167-9473(01)00065-2

69. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. 2016; 1–13. https://doi.org/10.1145/2939672.2939785

70. Venables WN, Ripley BD. Modern Applied Statistics with S. Fourth. New York: Springer; 2002.

71. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. Biometrics. 1988; 44: 837–845. https://doi.org/10.2307/2531595 PMID: 3203132

72. Youden WJ. Index for rating diagnostic tests. Cancer. 1950; 3: 32–35. https://doi.org/10.1002/1097-0142(1950)3:1<32::aid-cncr2820030106>3.0.co;2-3 PMID: 15405679

73. Perkins NJ, Schisterman EF. The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve. Am J Epidemiol. 2006; 163: 670–675. https://doi.org/10.1093/aje/kwj063 PMID: 16410346

74. Fawcett T. An introduction to ROC analysis. Pattern Recognit Lett. 2006; 27: 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

75. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez J-C, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12: 77. https://doi.org/10.1186/1471-2105-12-77 PMID: 21414208

76. R Core Team. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2019.

77. Donoho DL, Elad M. Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. Proc Natl Acad Sci. 2003; 100: 2197–2202. https://doi.org/10.1073/pnas.0437847100 PMID: 16576749

78. Donoho DL. For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. Commun Pure Appl Math. 2006; 59: 907–934. https://doi.org/10.1002/cpa.20131

79. Candes E, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Stat. 2007; 35: 2313–2351. https://doi.org/10.1214/009053606000001523

80. LeCun Y. Generalization and Network Design Strategies. 1989.

81. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998; 86: 2278–2324. https://doi.org/10.1109/5.726791

82. Simard P, LeCun Y, Denker JS. Efficient pattern recognition using a new transformation distance. NIPS'92: Proceedings of the 5th International Conference on Neural Information Processing Systems. 1992. pp. 50–58.

83. Mardia K V., Kent JT, Bibby JM. Multivariate Analysis. New York: Academic Press; 1979.

84. Hinton GE, Osindero S, Teh Y-W. A Fast Learning Algorithm for Deep Belief Nets. Neural Comput. 2006; 18: 1527–1554. https://doi.org/10.1162/neco.2006.18.7.1527 PMID: 16764513

85. Chawla N V., Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002; 16: 321–357. https://doi.org/10.1613/jair.953

86. Guyon I, Elisseeff A. An Introduction to Variable and Feature Selection. J Mach Learn Res. 2003; 3: 1157–1182.

87. Amershi S, Chickering M, Drucker SM, Lee B, Simard P, Suh J. ModelTracker: Redesigning Performance Analysis Tools for Machine Learning. Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems—CHI '15. New York, New York, USA: ACM Press; 2015. pp. 337–346. https://doi.org/10.1145/2702123.2702509

88. Newton KM, Peissig PL, Kho AN, Bielinski SJ, Berg RL, Choudhary V, et al. Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network. J Am Med Informatics Assoc. 2013; 20: e147–e154. https://doi.org/10.1136/amiajnl-2012-000896 PMID: 23531748

89. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, Ananthakrishnan AN, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ. 2015; 350: h1885. https://doi.org/10.1136/bmj.h1885 PMID: 25911572

90. Alzoubi H, Alzubi R, Ramzan N, West D, Al-Hadhrami T, Alazab M. A Review of Automatic Phenotyping Approaches using Electronic Health Records. Electronics. 2019; 8: 1235. https://doi.org/10.3390/electronics8111235