論文題目(Title of Thesis)　Claims-based algorithms for common chronic conditions were investigated using regularly collected data in Japan（日本の長期集積データを用いた主要な慢性疾患における Claims-based algorithms の検討）

氏　　名(Name)　原　湖楠

追加情報(Additional information)：

A follow-up paper based on the thesis has been published in
PLoS ONE 16(9): e0254394. https://doi.org/10.1371/journal.pone.0254394

博士論文

Claims-based algorithms for common chronic conditions were investigated using regularly collected data in Japan
（日本の長期集積データを用いた主要な慢性疾患における Claims-based algorithms の検討）

原　湖楠

Claims-based algorithms for common chronic conditions were investigated using regularly collected data in Japan
（日本の長期集積データを用いた主要な慢性疾患における Claims-based algorithms の検討）

東京大学大学院医学系研究科社会医学専攻
公衆衛生学教室

指導教員　小林　廉毅

原　湖楠

Contents

Summary

The literature of claims-based algorithm (CBA) has two features to be refined: the use of a chart review as a source of the gold standard; the procedure of searching for a fine-tuned CBA based on existing knowledge regarding target conditions. The first feature limits the population to which the CBA can be applied and the second makes the CBA construction procedure to be an overly complicated and cumbersome matter. Moreover, the burden of reviewing charts and searching for a fine-tuned CBA lead to a slow establishment of acceptable CBAs because it discourages researchers from CBA studies. The sluggish establishment of usable CBAs can be a big issue as the codes recorded in the claims for transmitting information about patients are supposed to change periodically. The dissertation focuses on CBAs for identifying patients with three common chronic medical conditions, hypertension, diabetes, and dyslipidemia, and (1) demonstrated the usefulness of health screening results as the source of gold standard (2) showed the power of statistical learning methods to develop an efficient CBA construction procedure; (3) proposed a course of action for an efficient CBA research. I believe that the series of techniques evaluated in the study should become essential in future CBA research.

List of Abbreviations

AUC	Area under the receiver operating characteristic curve
CART	Classification and regression tree
CBA	Claims-based algorithm
CI	Confidence interval
CV	Cross-validation
dBP	Diastolic blood pressure
EHR	Electronic health record
EPE	Expected prediction error
FBG	Fasting blood glucose
FDA	Flexible discriminant analysis
FY	Fiscal year
GAM	Generalized additive model
GBM	Gradient boosting machine
HbA1c	Hemoglobin A1c
HDL-C	High-density lipoprotein cholesterol
ICD-10	International Classification of Diseases and Related Health Problems, tenth revision
WHO-ATC	World Health Organization-anatomical therapeutic chemical
IDW	Inverse distance weighting
ISLE	Importance sampled learning ensemble
JMDC	Japan Medical Data Center
kNN	k-nearest neighbor

LDA	Linear discriminant analysis
LDL-C	Low-density lipoprotein cholesterol
NPV	Negative predictive value
OLS	Ordinary least squares
PDA	Penalized discriminant analysis
PPV	Positive predictive value
Prev.	Prevalence
RF	Random forest
ROC curve	Receiver operating characteristic curve
sBP	Systolic blood pressure
SD	Standard deviation
SGBM	Stochastic gradient boosting machine
Std.	Standardized
SVM	Support vector machine
TG	Triglyceride
VCM	Variance covariance matrix

Chapter 1

Introduction

1.1 Claims-based algorithm (CBA)

A growing body of research using medical and pharmacy claims data has been conducted in various fields including epidemiology, health service research, and health economics (Iizuka 2012; Einav, Finkelstein, and Schrimpf 2015; Schermerhorn et al. 2015; Abaluck and Gruber 2016; McWilliams et al. 2016; Nuti et al. 2016; Layton et al. 2017). Among them, a notable amount of research has used the claims data to assess medical conditions (Iizuka 2012; Einav, Finkelstein, and Schrimpf 2015; Nuti et al. 2016). Compared to other secondary data like electronic health records (EHRs), disease registries, and health screening results that are used to evaluate medical conditions, claims data occupies an important position in these research areas because of its completeness of the population coverage and relative easiness of long-term follow-up (Mitchell et al. 1994).

Nevertheless, claims data is subject to limitations due to potential imprecision in the identification of medical conditions (Virnig and McBean 2001; Taylor, Fillenbaum, and Ezell 2002; Rector et al. 2004; Kern et al. 2006; Klabunde, Harlan, and Warren 2006; Østbye et al. 2008). Because the claims are issued primarily for reimbursement to health care institutions, (1) information that is unnecessary for processing payments may not be collected or registered precisely in the claims forms; and (2) the diagnosis registered on claims may be relevant to testing for disease rather than to confirmed disease. The resulting misclassification of diagnosis can engender a substantial bias and undermine the credibility of the findings (Abrahamowicz et al. 2007). 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 (Quam et al. 1993; Hebert et al. 1999; Katz et al. 1997; Muhajarine et al. 1997; Robinson et al. 1997; Sands et al. 1999; Freeman et al. 2000; Andrade et al. 2002; Taylor, Fillenbaum, and Ezell 2002; Losina et al. 2003; Nattinger et al. 2004; Rector et al. 2004; Wilchesky, Tamblyn, and Huang 2004; Bullano et al. 2006; Gold and Do 2007; Nordstrom et al. 2007; Quan et al. 2009; Taylor et al. 2009; Cheng et al. 2011; Gorina and Kramarow 2011; Kawasumi et al. 2011; Scholes et al. 2011; Tu et al. 2011; Tessier-Sherman et al. 2013; Cheng et al. 2014; Chan et al. 2016; Walraven and Colman 2016; Yamana et al. 2016; Yamana et al. 2017; Hara et al. 2018). With the association measures attached to CBAs, researchers can assess the degree of uncertainty regarding their estimates of the effectiveness of their target treatments due to the potential imprecision of claims data. The information of the degree of uncertainty is particularly important when policymakers (or physicians) consider whether to adopt a policy (or treatment) which is evaluated using claims data because it leads to appropriate policy (or treatment) evaluation and application.

Since these 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 Japan, CBAs are not established for most of the medical conditions thus far. 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 (Van Walraven, Bennett, and Forster 2011). For this reason, researchers who are using claims data in Japan are facing an urgent need to establish CBAs for various medical conditions.

However, the literature of CBA still has two features to be refined: one regarding the source of the gold standard; another regarding the construction procedure of the CBA. In this study, I clarified obstacles in advancing research on CBA concerning these two features. I reviewed existing methods in the literature of CBA and made proposals on a better possible method that has not received much attention in the literature. I examined and discussed cases of three common chronic medical conditions, hypertension, diabetes, and dyslipidemia, about how these

proposals are considered superior in comparison with existing methods.

First, most previous studies reviewed medical charts to construct the gold standard for computing association measures of their CBA (Quam et al. 1993; Katz et al. 1997; Sands et al. 1999; Andrade et al. 2002; Losina et al. 2003; Wilchesky, Tamblyn, and Huang 2004; Bullano et al. 2006; Nordstrom et al. 2007; Quan et al. 2009; Cheng et al. 2011; Gorina and Kramarow 2011; Scholes et al. 2011; Tu et al. 2011; Cheng et al. 2014). Their results only apply to limited populations because the use of medical charts inevitably restricts the target population to those who visited clinics and hospitals on the review list. Therefore, it remains unclear if the CBA applies to a wider range of population and to what extent the claims data can gauge the number of patients with the target disease at the population level.

Routine health screening results can be a good substitute for medical charts especially for common chronic conditions which can be diagnosed with usual physical and laboratory examinations, e.g., hypertension, diabetes, and dyslipidemia. In Japan, under its universal health insurance system, some health insurance programs have established a system that collects medical and pharmacy claims data as well as annual health screening results regularly. This system provides us with the opportunity to assess the usability of CBAs to identify persons' medical conditions across a large and wide range of populations. Hara et al. (2018) demonstrated the usefulness of health screening results as the source of gold standard. There, they systematically and efficiently constructed an acceptable gold standard using health screening results and successfully assessed the usability of CBAs for hypertension, diabetes, and dyslipidemia with a large and wide range of populations. Here, I confirmed the results of Hara et al. (2018) with new data.

Second, previous studies have engaged in a knowledge-based condition-specific CBA construction procedure. When researchers have sought to find out a satisfactory CBA that identifies patients with their target medical condition, they needed to select input variables and decide how to incorporate variables in the CBA based on their experience or existing knowledge regarding the target condition.

For example, if one tries to obtain an acceptable CBA for identifying patients with hypertension, one may ask oneself the following questions: Which of using only the diagnostic codes

corresponding to hypertension, using only the medication codes corresponding to hypertension, or using both of them in the CBA is better?; Which codes of International Classification of Diseases and Related Health Problems, tenth revision (ICD-10)/World Health Organization-anatomical therapeutic chemical (WHO-ATC) should be designated as the diagnostic/medication codes corresponding to hypertension?; How many times should the diagnostic/medication codes appear in claims to consider a patient "test-positive" for hypertension? Researchers have assessed a large collection of knowledge-based candidate CBAs to select a fine-tuned CBA and iterated the procedure if there are multiple target conditions. I illustrated how complicated and cumbersome the procedure of knowledge-based condition-specific CBA construction is in the dissertation (subsection 2.5.1). A method that fine-tunes CBAs regardless of the level of knowledge for the target condition and without condition-specific modifications of the procedure can refine the procedure to be smarter and more convenient. Nonetheless, such a method has not yet been established.

To this end, it is natural to think of the usage of regression methods as in some previous CBA research (Muhajarine et al. 1997; Freeman et al. 2000; Taylor, Fillenbaum, and Ezell 2002; Nattinger et al. 2004; Gold and Do 2007; Østbye et al. 2008; Quan et al. 2009; Kawasumi et al. 2011). Since it is known that regression methods often work poorly in the accuracy of prediction when the number of input variables is large relative to the sample size (Zou and Hastie 2005), input variables may need to be selected before implementing a regression method to obtain a satisfactory CBA. 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.

Statistical learning methods, which are overviewed in the next section, are promising technologies to overcome the problem of regression methods. In the dissertation, I define statistical learning methods as the methods that aim to minimize the estimator of the risk functional via the hyperparameter tuning and regard them as a subset of machine learning methods. Machine learning methods can be broadly defined as the computational methods that use existing knowledge to improve the performance of their prediction (Mohri, Rostamizadeh, and Talwalkar 2012). They can rely on investigated facts and known features specific to the subject to which

the method will be applied besides the risk functional criterion. Statistical learning methods are more mathematically and statistically tractable compared to the other groups of machine learning methods, and this tractability is the reason why I focused on statistical learning methods. Note that, in general, there seems to be only a vague boundary of the statistical learning domain in the machine learning world as it evolves according to the times.

A few researchers have attempted to use statistical learning methods in the context of CBA (Sands et al. 1999; Nordstrom et al. 2007; Scholes et al. 2011; Chan et al. 2016; Walraven and Colman 2016). However, they did not try to circumvent the knowledge-based condition-specific CBA construction procedure. I applied statistical learning methods to a dataset that input variables were chosen to be common to all target conditions; the dataset consists of age, gender, and all ICD-10/WHO-ATC codes with a letter followed by two digits as input variables. This simple device renders the procedure to be condition-invariant.

Additionally, previous studies only used a specific statistical learning method without sufficient support. Because a statistical learning method that suits the context is yet unknown, it is important to explore which statistical learning method suits for the CBA setting. As a starting point for the development of CBA construction procedure using statistical learning methods, I investigated popular statistical learning methods with the theories behind them to examine the outline of what kind of method seems to work in the context of CBA. Although one can think of a method aiming at the further improvement of prediction accuracy than the methods used in this study according to the context of machine learning, I believe that the findings obtained from this study are still valuable as a place to begin the discussion on the development of efficient CBA construction procedure.

Third, the burden of reviewing charts and searching for a fine-tuned CBA lead to a slow establishment of acceptable CBAs because it discourages researchers from CBA studies. The slow establishment can be a big issue when the codes recorded in the claims for transmitting information about patients are supposed to change periodically. As a result of a coding scheme change, re-construction and re-assessment of CBAs may be necessary, and if CBA studies only proceed gradually, the scheme change should cause a huge challenge in the continuous usage of administrative data. 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 (Khera, Dorsey, and Krumholz 2018).

The methods discussed in this study can sidestep these obstacles and may boost the implementation of CBA research. Chart reviewing can be avoided by the use of regularly collected data like annual health screening results. EHRs and disease registries are other possible candidates in this direction. Fine-tuned CBAs can be efficiently searched by the use of a condition-invariant procedure in the CBA construction. Researchers can uniformly apply the procedure to construct a CBA for each of their target conditions and compare it against their gold standard that is constructed from the regularly collected data. This course of action should greatly reduce the burden of CBA research, and thereby strongly supports the seamless usage of administrative data in an environment of a periodic coding scheme change.

CBA research has a closely related research area called "phenotyping" (Newton et al. 2013). Phenotyping algorithms aim to identify medical conditions (or phenomic traits) like CBA but with EHRs besides claims data. Regardless of the similarity, there are two large differences between these research areas: (1) phenotyping algorithms are developed assuming that they will be mostly applied to genomics studies that require more stringent accuracy than the fields to which CBAs are assumed to be applied; (2) phenotyping algorithms are based on much more comprehensive and complicated information than the information on which CBAs are based. Because of these differences, it will be very difficult to adapt the concepts implied in this study. Nevertheless, as phenotyping algorithms can be satisfactorily accurate to be used in gold standard construction of CBA research, they can aid the implementation of CBA research when EHRs are available.

To recapitulate, the dissertation focuses on CBAs for identifying patients with three common chronic medical conditions, hypertension, diabetes, and dyslipidemia, and (1) demonstrated the usefulness of health screening results as the source of gold standard following Hara et al. (2018); (2) showed the power of statistical learning methods to develop an efficient CBA construction procedure; (3) proposed a course of action for an efficient CBA research.

1.2 An overview of statistical learning methods

Statistical learning methods aim to predict the outcome for new observation using the data at hand. To understand the relationship between regression methods (e.g., linear regression, logistic regression, and alike) and statistical learning methods, I state the learning problem summarized in Vapnik (1999).

Some generic notations are necessary to be defined before the explanation. Denote f as a generic notation for a probability density function, e.g., $f_X(X = x_0)$. When f indicates the whole distribution like $f_X(X)$, and the subscript of f and the object in the parenthesis are the same, I omit the subscript, i.e., $f_X(X) = f(X)$. The expectation of a function $g(X)$ taken over a distribution $f(X)$ is defined as

$$E_{f(X)}[g(X)] \equiv \int g(X)f(X)dX.$$

I use $E_X[g(X)]$ (the expectation of $g(X)$ taken over a distribution of X) or $E[g(X)]$ (the expectation of $g(X)$ taken over a distribution) when the distribution taking over the expectation is clear from the context. Similarly, let $\mathrm{Var}_{f(X)}[\cdot]$ indicate that the expectation in the variance formula is taken over a distribution $f(X)$, and $\mathrm{Var}_X[\cdot]$ and $\mathrm{Var}[\cdot]$ be the short form of it.

The model of learning or estimating the underlying function of an outcome from a sample dataset, which is drawn from the target population, is described by five components. At first, a set of random vectors $X \in \mathcal{X} \subset \mathbb{R}^p$ which are drawn independently from the p-dimensional input distribution of the target population, $f(X)$. I implicitly include a constant as the first element of X for notational simplicity.

Second, a *supervisor* that returns an output $Y \in \mathcal{Y} \subset \mathbb{R}$ for every input vector X according to the conditional distribution of the output of the target population $f(Y|X)$. Whether the output type is continuous or discrete affects the representation of the function which we seek to learn. This distinction in output type has led to a naming convention for the prediction tasks: *regression* when we predict continuous outputs, and *classification* when we predict discrete outputs. For K-class classification ($K \geq 2$), I use a set $\{0, 1, \ldots, k, \ldots, K-1\}$ as a notation for K classes except when stated otherwise.

Third, a sample dataset $\mathcal{T} \equiv \{(x_1, y_1), \ldots, (x_N, y_N)\}$ which is assumed to be N independent identically distributed random observations drawn from the joint distribution of X and Y, $f(X, Y) = f(X)f(Y|X)$.

Fourth, a *learning machine* which is capable of implementing a set of candidate functions. Here, it is convenient to consider regression and classification separately. For regression, we want to find a function $a : \mathcal{X} \to \mathcal{Y}$ such that $a(X)$ approximates Y. Thus, a set of candidate functions suitable for regression is $\{a(X; \theta) : \theta \in \Theta\}$, where functions are characterized by a vector of parameters θ in a parameter space Θ. For instance, candidate functions can be specified as a simple linear in parameters model with an arbitrary p-dimensional parameter space: $\{X^T\theta : \theta \in \Theta \subset \mathbb{R}^p\}$.

By contrast, for K-class classification, although $a(X)$ is an interest as well, frequently, a function $b : \mathcal{X} \to \mathbb{R}^K$ such that $b(X) = (b_0(X), b_1(X), \ldots, b_k(X), \ldots, b_{K-1}(X))^T$ is a vector of scores of the propensity for the assignment to classes attracts more attention than it. The domain of the function is typically a subset of $\mathbb{R}^K$. For example, researchers often seek $b : \mathcal{X} \to [0, 1]^K$ such that $\sum_{k=0}^{K-1} b_k(X) = 1$ and $b_k(X)$ approximates the conditional probability of the assignment to class k, $\Pr(Y = k|X)$, because it is convenient for the interpretation of the results.

Another particular example is a prediction function for two-class classification. With only two classes, a score of the propensity for the assignment to either of the classes is sufficient for the prediction purpose. Consequently, one of the elements of $b(X)$, say, $b_1 : \mathcal{X} \to \mathbb{R}$ such that $b_1(X)$ is a score of the propensity for the assignment to class 1, is often a primary interest of researchers.

Therefore, a set of candidate functions suitable for classification can be either of $\{a(X; \theta) : \theta \in \Theta\}$ or $\{b(X; \theta) : \theta \in \Theta\}$. Note that even when the objective of classification is to find $a(X)$, this is usually carried out taking

$$a(X) = \arg\max_k b_k(X)$$

after the estimation of $b(X)$. For brevity of explanation, the response functions $a(X; \theta)$ and $b(X; \theta)$ will be collectively referred to as $\Psi_\theta(X)$ when the distinction between them is unnecessary.

The final component is a measure of the loss or discrepancy, a loss function $L(Y, \Psi_\theta(X))$, between the response Y of the supervisor and the response $\Psi_\theta(X)$ provided by the learning machine for a given X.

The problem of learning is that of choosing the function $\Psi_\theta(X)$ which is the best available approximation to the supervisor's response in terms of the loss function $L(Y, \Psi_\theta(X))$ from the given set of functions $\{\Psi_\theta; \theta \in \Theta\}$ based on the sample dataset $\mathcal{T}$. Consider the expected value of the loss, given by the *risk functional* (Vapnik 1999):

$$R(\theta) \equiv E_{X,Y}[L(Y, \Psi_\theta(X))].$$

The popular loss function is a squared error loss for regression and a negative log-likelihood or *log-loss* for classification. For sample (x_i, y_i), the squared error loss is defined as

$$L(y_i, a(x_i; \theta)) = \{y_i - a(x_i; \theta)\}^2,$$

and the log-loss as

$$L(y_i, b(x_i; \theta)) = -\log b_{y_i}(x_i; \theta),$$

where the log-loss implicitly assumes $\sum_{k=0}^{K-1} b_k(X) = 1$. The risk functional with the squared error loss is called *expected prediction error* (EPE):

$$\text{EPE}(\theta) \equiv E_{X,Y}[\{Y - a(X; \theta)\}^2].$$

Now, the goal of the statistical learning can be summarized as to find the function $\Psi_{\theta_0}, \theta_0 \in \Theta$ which minimizes the risk functional $R(\theta)$ over the class of functions $\{\Psi_\theta : \theta \in \Theta\}$ when the only available information is the sample dataset $\mathcal{T}$.

Regression methods can be interpreted as methods of finding θ that minimize the empirical risk functional $R(\theta)$:

$$\underset{\theta}{\text{argmin}} \frac{1}{N} \sum_{i=1}^{N} L(y_i, \Psi_\theta(x_i)).$$

Linear regression or ordinary least squares (OLS) specifies the output function $a(X; \theta)$ as linear

in parameters and the loss function as squared error loss:

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} (y_i - x_i^T \theta)^2.$$

Binary logistic regression specifies the output function $b(X; \theta)$ in the logit form and the loss function as the log-loss:

$$\underset{\theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \{-\log b_{y_i}(x_i; \theta)\},$$

where

$$b_{y_i}(x_i; \theta) = \frac{y_i \exp x_i^T \theta}{1 + \exp x_i^T \theta} + \frac{1 - y_i}{1 + \exp x_i^T \theta}.$$

Regression methods are known to provide $\sqrt{N}$-consistent estimators under certain regularity conditions. There are many admirable textbook treatments with regard to regression methods and asymmptotic analysis of them (e.g., Wooldridge (2010) and Greene (2012)).

In the case of a fixed number of inputs p with sample size $N \to \infty$ or sufficiently large relative to p, regression methods work perfectly. Regression methods provide a consistent estimator with a reasonable sampling variance, and consequently, provide a useful predictive value at target value $X = x_0$ by just plugging in the values in the estimated function. However, when p is large relative to N including the case of $p > N$, it is well known that regression methods often work poorly in the interpretation of the estimated parameters and the accuracy of prediction on future data (Zou and Hastie 2005). Since the regularity condition no longer holds when $p > N$, regression methods fail to maintain the consistency, and hence, any interpretation is left on the estimator. Moreover, even if $p \leq N$, one can hardly acquire an adequately small sampling variance to gain some meaningful insights from the results when p is large relative to N.

In contrast to regression methods, statistical learning methods pursue to minimize the risk functional $R(\theta)$ more directly using the *hyperparameter* rather than relying on the sample analogue of the risk functional. The hyperparameter aids the model to attain the minimum risk functional in the following way. Statistical learning methods randomly divide the sample dataset into two parts: a *training set* and a *validation set*. For each candidate value of the hyperparameter, an estimation of the parameter of the model is conducted with the training set,

and an estimator for the risk functional of the estimated model is computed by the average loss of the model in the validation set. The hyperparameter is subsequently tuned to be the value that minimizes the estimator of the risk functional.

A method called *cross-validation* (CV) is also used to estimate the risk functional. Although CV is computationally much harder than the calculation of the average loss in the validation set, CV is a more efficient way of estimating the risk functional.

For two-class classification, Bradley (1997) proposed a method based on one minus the area under the receiver operating characteristic curve (AUC) instead of the risk functional for the hyperparameter tuning. As the calculation of the AUC only requires a ranked list of samples of their propensity for the assignment to classes, the method is robust to the monotonic transformation of the functional form of the prediction function. Although the one minus AUC approach is not covered by the risk functional approach of Vapnik (1999), the notion of the risk functional approach is later refined to include such approach (Chen et al. 2009). There, the risk functional approach of Vapnik (1999) is subsumed as the *pointwise approach*, and the one minus AUC approach is categorized as the *listwise approach*. However, I continue to assume the word "risk functional" to mean the risk functional of Vapnik (1999) because the formal definition of the refinement involves complicated and confusing notions regarding the distribution of the expected value of the loss to be taken. All of the following statement regarding the risk functional can be extended to the one minus AUC approach without any modification.

A typical hyperparameter of statistical learning methods is the coefficient for the *regularization* term. Suppose there are a dataset generated from an underlying function with some form of error and a set of functions that are candidates for the best underlying function approximation. The regularization principle, which is first introduced by Tikhonov (1963), imposes some form of smoothness constraints on the candidate functions to find the best approximating function given the dataset. Adapting the regularization technique to the statistical learning setting yields a general class of regularization problems:

$$\min_{\theta} \left\{ \sum_{i=1}^{N} L(y_i, \Psi_\theta(x_i)) + \lambda J(\theta) \right\}, \tag{1.1}$$

where $\lambda \geq 0$ is a regularization coefficient and $J(\cdot)$ is the regularization term. Commonly used

regularization term is an L_2-*penalty*

$$J(\theta) = \|\theta\|_2^2 = \sum_{j=1}^{p} \theta_j^2,$$

and an L_1-*penalty*

$$J(\theta) = \|\theta\|_1 = \sum_{j=1}^{p} |\theta_j|.$$

This formulation emcompasses most of statistical learning methods covered here.

Although it is better to tune the hyperparameter to minimize the estimator of the risk functional, it is often prespecified based on existing knowledge to avoid high computational burden. As which of the whole sample dataset and the training set should be used for estimation of parameters depends on whether the hyperparameter is prespecified or not, I do not distinguish between the sample dataset and the training set in the following exposition of the estimation detail of statistical learning methods. There, I use the word "dataset" with a standard sample size notation N.

The distinction between the regression and the statistical learning is further clarified through the following example. Suppose the sample dataset $\mathcal{T}$ arises from a linear model

$$Y = X^T\theta_0 + \epsilon \tag{1.2}$$

with $\theta_0 \in \Theta \subset \mathbb{R}^p$, $E(\epsilon|X) = 0$, and $\mathrm{Var}(\epsilon|X) = \sigma^2 < \infty$. Let the estimator from a linear regression be $\hat{\theta}$. Then, the EPE at the target input x_0 can be decomposed into three elements:

$$\begin{aligned}
\mathrm{EPE}(\theta|X = x_0) &= E_{\mathcal{T}}\left[E_{Y|X=x_0}[\{Y - x_0^T\hat{\theta}\}^2|X = x_0]\right] \\
&= E_{Y|X=x_0}[\{Y - x_0^T\theta_0\}^2|X = x_0] \\
&\quad +\{x_0^T\theta_0 - E_{\mathcal{T}}[x_0^T\hat{\theta}]\}^2 \\
&\quad +E_{\mathcal{T}}[\{E_{\mathcal{T}}[x_0^T\hat{\theta}] - x_0^T\hat{\theta}\}^2] \\
&= \sigma^2 + \mathrm{Bias}^2[x_0^T\hat{\theta}] + \mathrm{Var}_{\mathcal{T}}[x_0^T\hat{\theta}],
\end{aligned}$$

where $E_{\mathcal{T}}[\cdot]$ indicates that the expectation is taken over the sampling distribution of the sample

dataset $\mathcal{T}$. Since the first term σ^2 cannot be reduced by devising the estimation method, minimizing the EPE is the same as minimizing the sum of squared bias and the sampling variance. Under the linear model (1.2), the unbiasedness of OLS assures the unbiasedness of the resulting prediction as well:

$$E_{\mathcal{T}}[x_0^T\hat{\theta}] = x_0^T E_{\mathcal{T}}[\hat{\theta}] = x_0^T\theta_0 \Leftrightarrow \text{Bias}[x_0^T\hat{\theta}] = 0.$$

Thus, we now know that the linear regression produces a least bias estimator for the prediction. But still, some methods may achieve their better prediction performance through a bias-variance trade-off. Statistical learning methods are such methods that aim to minimize EPE by the reduction of sampling variance along with paying the cost of some bias.

The remaining of this section is largely based on Hastie, Tibshirani, and Friedman (2009) and overviews popular statistical learning methods: (1) Discriminant analysis; (2) Generalized additive model; (3) k-nearest neighbor; (4) Support vector machine; (5) Penalized regression; (6) Tree-based model; (7) Neural network.

1.2.1 Discriminant analysis

The *linear discriminant analysis* (LDA), which is originally proposed by Fisher (1936), aims to discriminate between two or more classes with a linear discriminant function that maximizes the ratio of the between-class variance to the within-class variance. This subsection only deals with K-class classification, on which the discriminant analysis mainly focuses.

Denote the between-class variance covariance matrix (VCM) of the input vector X as $B = \text{Var}_Y(E[X|Y])$ and the class k's within-class VCM of X as $W_k = \text{Var}(X|Y = k)$. Now, consider a linear discriminant function $Z = \nu^T X$, $\nu \in \mathbb{R}^p$. The between-class variance of Z is

$$\text{Var}_Y(E[Z|Y]) = \text{Var}_Y(\nu^T E[X|Y]) = \nu^T \text{Var}_Y(E[X|Y])\nu = \nu^T B\nu,$$

and the class k's within-class VCM of Z is

$$\text{Var}(Z|Y = k) = \nu^T \text{Var}(X|Y = k)\nu = \nu^T \text{Var}(X|Y = k)\nu = \nu^T W_k \nu.$$

The LDA assumes a common VCM, W, for all classes, $\forall k$, $W_k = W$.

Then, the Fisher's problem amounts to finding ν that maximizes the ratio of the between-class variance of Z to the within-class variance of Z:

$$\arg\max_{\nu} \frac{\nu^T B \nu}{\nu^T W \nu}.$$

The solution, ν_1, is shown to be $W^{-1/2}$ multiplied by the eigenvector of the largest eigenvalue of $W^{-1/2} B W^{-1/2}$. Similarly one can find the second best linear discriminant function $\nu_2^T X$ by finding ν orthogonal to ν_1 that maximizes $\nu^T B \nu / \nu^T W \nu$: the solution ν_2 is $W^{-1/2}$ multiplied by the eigenvector of the second largest eigenvalue of $W^{-1/2} B W^{-1/2}$. And this procedure can be continued L times to find a sequence of discriminant coordinates $\{\nu_l\}_{l=1}^{L}$. As we need at most $K-1$ discriminant functions to separate input space into K classes, $L \leq K-1$. In practice, discriminant coordinates are estimated using the estimator of B and W, e.g.,

$$\hat{B} = \frac{1}{N} \sum_{k=0}^{K-1} |\mathcal{S}_k| (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^T \quad \text{and} \quad \hat{W} = \frac{1}{N} \sum_{k=0}^{K-1} \sum_{i \in \mathcal{S}_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)^T,$$

where $\mathcal{S}_k = \{i : y_i = k\}$, $|\mathcal{S}_k|$ is the cardinality of $\mathcal{S}_k$, $\bar{x}_k = \sum_{i \in \mathcal{S}_k} x_i / |\mathcal{S}_k|$, and $\bar{x} = \sum_{i=1}^{N} x_i / N$.

As L is at most $K-1$, the LDA usually achieves a considerable dimension reduction compared to the p-dimensional input space and may help us to understand informative attributes of the data. Besides, the discriminant functions can be interpreted as a score of the propensity for a certain class assignment and this interpretation leads to a further generalization of the LDA.

Consider an optimal scoring problem that transforms class labels to scores which are optimally predicted by linear regression on X. Suppose $s : \mathcal{Y} \to \mathbb{R}^L$ is a function that assigns $L \leq K-1$ scores to the classes. Then the optimal scoring problem solves the following minimization problem:

$$\operatorname*{argmin}_{s, \{\beta_l\}_{l=1}^{L}} \sum_{l=1}^{L} \sum_{i=1}^{N} \{s_l(y_i) - x_i^T \beta_l\}^2,$$

where s_l is the lth component of s. It can be shown that the sequence of discriminant coordinates $\{\nu_l\}_{l=1}^{L}$ derived from the Fisher's problem is identical to the sequence $\{\beta_l\}_{l=1}^{L}$ up to a constant (Mardia, Kent, and Bibby 1979).

From this re-formalization of the LDA, one can think of a basis expansion and an application of regularization principle to generalize the LDA:

$$\underset{s,\{\beta_l\}_{l=1}^{L}}{\operatorname{argmin}} \sum_{l=1}^{L} \left[\sum_{i=1}^{N} \{s_l(y_i) - h(x_i)^T \beta_l\}^2 + \lambda J(\beta_l) \right],$$

where $h(\cdot)$ is a flexible transformation function. This generalization of the LDA is proposed by Hastie, Tibshirani, and Buja (1994) as the *flexible discriminant analysis* (FDA), and by Hastie, Buja, and Tibshirani (1995) as the *penalized discriminant analysis* (PDA). The FDA and PDA successfully incorporate nonlinearity while maintaining the dimension reduction aspect of the LDA.

Although the discriminant analysis is good at classification in a moderate number of inputs, they are poor at dealing with sparse high-dimensional inputs (i.e., most of the entries of the inputs are zero for each observation, and the number of the inputs is large relative to the sample size). Because the model underlying the discriminant analysis assumes that the within-class distribution of every class is nondegenerate for all inputs, inputs which have the same value for all samples in some class do not fit for the model. Consequently, computer programs that execute the discriminant analysis usually refuse a dataset with such inputs, and we need to discard those inputs before the analysis if we seek to apply the discriminant analysis to the dataset. One can think of an *unsupervised learning* method that extracts essential components of the inputs, e.g., *principal component analysis* (Mardia, Kent, and Bibby 1979), before the analysis to alleviate the problem. Nevertheless, there is no standard way of an input pre-processing in this direction for the discriminant analysis yet.

1.2.2 Generalized additive model

A linear model fails to capture nonlinear effects, and this fact distorts the ability to predict outcomes by it. Although one can think of adding nonlinear terms to the linear model by hand, one never knows whether the additional terms are sufficient for the model or not. Hastie and Tibshirani (1986) proposed the *generalized additive model* (GAM), in which flexible nonlinear effects can be incorporated automatically.

For regression, a candidate function $a(X;\theta)$ in the GAM is specified as

$$\text{link}[a(X;\theta)] = \sum_{j=1}^{p} \alpha_j(X_j;\theta_j),$$

where link$[\cdot]$ is a *link function*, X_j is the jth input, and θ_j is the jth set of parameters. When the link is the *identity link*, i.e., link$[a] = a$, and the loss function is the squared error loss, the model is an extension of the linear regression, and it is called the *additive linear regression model*.

For two-class classification, candidate functions $b(X;\theta)$ in the GAM are specified as

$$\text{link}[b_1(X;\theta)] = \sum_{j=1}^{p} \alpha_j(X_j;\theta_j).$$

This concept can be extended to K-class classification. When the link is the *logit link*, link$[b]$ = logit(b), and the loss function is the log-loss, the model is an extension of the logistic regression, and it is called the *additive logistic regression model*.

Each function $\alpha_j(X_j;\theta_j)$ is fitted by a *scatterplot smoother* (e.g., a *cubic smoothing spline* or *kernel smoother*) using the *backfitting algorithm*, where the degrees of freedom for the smoothers is a hyperparameter of the model. A detailed description of estimation methods is covered in Hastie and Tibshirani (1990).

The additive linear regression model with a cubic smoothing spline is shown to be the minimizer of

$$\sum_{i=1}^{N}\{y_i - \sum_{j=1}^{p} \alpha_j(x_{ij};\theta_j)\}^2 + \sum_{j=1}^{p} \lambda_j \int \{\alpha_j^{''}(t_j)\}^2 dt_j,$$

where x_{ij} is the jth input of the ith sample (Hastie, Tibshirani, and Friedman 2009). Thus, the GAM can be interpreted as a regularization problem (1.1). The GAM is highly flexible in incorporating nonlinearity of a moderate number of inputs, however, the difficulty in the hyperparameter tuning and the need of input pre-processing if the number of inputs is large narrow the area of suitable application of the model. Although a progress has been made in the smoothing parameter tuning and the automatic variable selection in a sparse high-dimensional setting recently (Lin and Zhang 2006; Ravikumar et al. 2009), these potentially innovative methods are still computationally prohibitive for large scale data.

1.2.3 k-nearest neighbor

The *k-nearest neighbor* (kNN) classifier appears as a natural estimator for the *nonparametric discriminatory analysis* (Fix and Hodges 1951). As the kNN is mainly applied to classification, this subsection concentrates on classification. To avoid a notational confusion caused by the usage of the same letter 'k' in K-class classification and the k-nearest neighbor, let K-class be J-class instead in this subsection: $\mathcal{Y} = \{0, 1, \ldots, j, \ldots, J-1\}$. k is a hyperparameter of the model.

Let the distance of a query point x_0 and a sample input x_i can be measured by a designated distance metric $D(x_0, x_i)$. In the kNN, we first find a set of indices of k training samples, $\mathcal{S}_{k(x_0)} \subset \{1, 2, \ldots, N\}$, which are the k closest neighbors to the query point. Then, we predict a class probability of the query point from the frequency of the class of the k-nearest neighbors (*voting* estimator, Fix and Hodges (1951)),

$$b_j(x_0; \theta) = \frac{1}{k} \sum_{i \in \mathcal{S}_{k(x_0)}} I(y_i = j),$$

or the inverse distance weighted frequency of the class of the k-nearest neighbors (*inverse distance weighting* (IDW) estimator, Shepard (1968)),

$$b_j(x_0; \theta) = \frac{\sum_{i \in \mathcal{S}_{k(x_0)}} w(x_0, x_i) I(y_i = j)}{\sum_{i \in \mathcal{S}_{k(x_0)}} w(x_0, x_i)}, \quad \text{where} \quad w(x_0, x_i) = \frac{1}{D(x_0, x_i)}.$$

Typically, the Euclidean (L_2) distance is used as the distance metric, i.e., $D(x_0, x_i) = \|x_0 - x_i\|_2$. Besides, inputs may be better to be standardized to have mean zero and variance one when units of the inputs are different from each other. The distance metric can be a more general distance measure like the *L_p-distance* or the *Mahalanobis distance*, and in the case of categorical inputs, a distance measure like the *Hamming distance* may be more suitable for a distance metric. Designing the distance metric in the kNN is difficult as an appropriate distance metric depends on the setting.

Although the kNN creates a highly flexible nonparametric estimator for classification, the lack of the interpretability of the method discourages the use of the method in biomedical and

clinical research except research related to the field of image recognition, where the kNN had established an era by the invention of the *tangent distance* (Simard, LeCun, and Denker 1992).

1.2.4 Support vector machine

The *support vector machine* (SVM) is developed as an extension of the *optimal separating hyperplane* (Boser, Guyon, and Vapnik 1992; Cortes and Vapnik 1995). The idea is to map the inputs into a high-dimensional input space through some prespecified nonlinear mapping and to construct an optimal separating hyperplane in the enlarged space. Although the SVM can be extended to K-class classification or regression, the main focus of the model is on two-class classification, with which I deal here. In the SVM literature, output space is usually defined as $\mathcal{Y} = \{-1, 1\}$, and I follow this convention.

First, consider an optimal separating hyperplane between two perfectly separable classes (Rosenblatt 1958). Define a hyperplane L by $\{X : X^T\theta = 0, \|\theta\|_2 = 1\}$, where $\|\theta\|_2 = 1$ is for a normalization, and a classification rule by

$$a(X; \theta) = \text{sign}[X^T\theta].$$

As the signed distance from a point X to the hyperplane L is $X^T\theta$, the optimal separating hyperplane that creates the biggest margin between the training points for two classes satisfies

$$\underset{\{\theta:\|\theta\|_2=1\}}{\arg\max}\ M \ \text{ s.t. } \forall i,\ \ y_i(x_i^T\theta) \geq M,$$

which is equivalent to

$$\underset{\theta}{\text{argmin}}\ \|\theta\|_2 \ \text{ s.t. } \forall i,\ \ y_i(x_i^T\theta) \geq 1.$$

The latter formulation is more convenient to solve as the norm constraint on θ is dropped and M is eliminated. Now, the margin is formally defined as the area that satisfies $|y_i(x_i^T\theta)| \leq M$ under the normalization constraint $\|\theta\|_2 = 1$ and $|y_i(x_i^T\theta)| \leq 1$ otherwise.

Next, suppose that the classes are linearly nonseparable. The optimal separating hyperplane is redefined as the hyperplane that maximizes the margin, but allows for some points to be on the

wrong side of the boundary of the margin subject to a given upper bound of the total proportional amount of the slack, C:

$$\underset{\{\theta:\|\theta\|_2=1\}}{\arg\max}\ M \ \text{ s.t. } \begin{cases} \forall i,\ \ y_i(x_i^T\theta) \geq M(1-\xi_i) \ \text{ and } \ \xi_i \geq 0, \\ \sum_{i=1}^N \xi_i \leq C \end{cases},$$

which is equivalent to

$$\underset{\theta}{\operatorname{argmin}} \|\theta\|_2 \ \text{ s.t. } \begin{cases} \forall i,\ \ y_i(x_i^T\theta) \geq 1-\xi_i \ \text{ and } \ \xi_i \geq 0, \\ \sum_{i=1}^N \xi_i \leq C \end{cases}. \tag{1.3}$$

Moreover, the solution of (1.3) can be shown to be equivalent to that of

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^N \max(1-y_i(x_i^T\theta), 0) + \lambda\|\theta\|_2^2,$$

where λ is the hyperparameter reflecting the cost of the slack. The loss function $L(Y, X^T\theta) = \max(1 - Y(X^T\theta), 0)$ is known as the *hinge loss*. It is clear that the problem also forms a class of regularization problems (1.1). Given the hyperparameter λ, the solution is

$$\hat{\theta}_\lambda = \frac{1}{2\lambda}\sum_{i=1}^N \hat{\alpha}_i y_i x_i,$$

where

$$\hat{\alpha}_i = \begin{cases} 0 & \text{if } \ y_i(x_i^T\hat{\theta}_\lambda) > 1 \ \text{ (samples correctly classified and outside the margin),} \\ [0,1] & \text{if } \ y_i(x_i^T\hat{\theta}_\lambda) = 1 \ \text{ (samples sitting on the boundary of the margin),} \\ 1 & \text{if } \ y_i(x_i^T\hat{\theta}_\lambda) < 1 \ \text{ (samples inside the margin or wrongly classified).} \end{cases}$$

Therefore, the estimated optimal separating hyperplane at λ can be written as

$$\begin{aligned} & X^T\hat{\theta}_\lambda = 0 \\ \Leftrightarrow \quad & \frac{1}{2\lambda}\sum_{i=1}^{N}\hat{\alpha}_i y_i X^T x_i = 0. \end{aligned}$$

From the formula, you can see that the samples correctly classified and outside the margin do not contribute to the prediction rule.

Finally, to enlarge the input space, substitute the transformed input vectors $h(x_i)$ for the raw input vectors x_i:

$$\underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{N} \max(1 - y_i(h(x_i)^T\theta), 0) + \lambda\|\theta\|_2^2.$$

Then, the estimated generalized hyperplane at λ is

$$\begin{aligned} & h(X)^T\hat{\theta}_\lambda = 0 \\ \Leftrightarrow \quad & \frac{1}{2\lambda}\sum_{i=1}^{N}\hat{\alpha}_i y_i h(X)^T h(x_i) = 0 \\ \Leftrightarrow \quad & \frac{1}{2\lambda}\sum_{i=1}^{N}\hat{\alpha}_i y_i K(X, x_i) = 0, \end{aligned}$$

where $K(X, X') \equiv h(X)^T h(X')$ is a type of function known as the *kernel function*, and the resulting classification rule is

$$a(X; \hat{\theta}_\lambda) = \text{sign}[\frac{1}{2\lambda}\sum_{i=1}^{N}\hat{\alpha}_i y_i K(X, x_i)].$$

It is known that the sufficient knowledge for the estimation of the SVM is the kernel function, and the input transforming function $h(X)$ is not necessarily required. The popular choices for

the kernel function in the SVM literature are

$$\begin{aligned}
\text{Linear: } K(X, X') &= X^T X', \\
d\text{th-degree polynomial: } K(X, X') &= (1 + X^T X')^d, \\
\text{Radial basis: } K(X, X') &= \exp(-\gamma \|X - X'\|_2^2), \\
\text{Sigmoid: } K(X, X') &= \tanh(\kappa_1 X^T X' + \kappa_2).
\end{aligned}$$

There are two features of the hinge loss that make the SVM robust to outliers. First, under the hinge loss, samples correctly classified and outside the margin do not contribute to the prediction rule. Second, the hinge loss gives a linear penalty rather than a quadratic penalty to samples inside the margin or wrongly classified. Nonetheless, a squared hinge loss that gives a quadratic penalty to the samples may improve the performance when the effect of the outliers is negligible.

The SVM is extremely computationally intensive for the large sample size. Researchers have developed an efficient algorithm that is computationally feasible for the linear kernel SVM (Rong-En et al. 2008). However, it is still challenging to employ other kernel functions in the large sample size setting.

1.2.5 Penalized regression

As a natural consequence of the development of regularization techniques, researchers have applied the regularization principle to the linear regression and created a *penalized least squares* estimator:

$$\underset{\theta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} (y_i - x_i^T \theta)^2 + \lambda J(\theta) \right\}.$$

Penalized least squares methods using the L_2-penalty and the L_1-penalty as the regularization term are called *ridge regression* (Hoerl and Kennard 1970) and the *lasso* (Tibshirani 1996), respectively.

When the inputs are mutually independent and standardized to have mean zero and variance one, the estimator of the coefficient of the jth input of the ridge $\hat{\theta}_j^{Ridge}$ and that of the lasso $\hat{\theta}_j^{Lasso}$

can be expressed by the corresponding OLS estimator $\hat{\theta}_j^{OLS}$ and the regularization coefficient λ:

$$\begin{cases} \hat{\theta}_j^{Ridge} = \frac{\hat{\theta}_j^{OLS}}{1+\lambda}; \\ \hat{\theta}_j^{Lasso} = \text{sign}(\hat{\theta}_j^{OLS})\{\max(|\hat{\theta}_j^{OLS}| - \lambda, 0)\}. \end{cases}$$

The ridge regression does a proportional shrinkage, while the lasso shifts each OLS estimator by a constant factor λ, truncating at zero. The ridge regression cannot produce a parsimonious model as it keeps all inputs in the model in the same way as the linear regression. On the other hand, the lasso does both continuous shrinkage and automatic variable selection to obtain a parsimonious model. The automatic variable selection is an attractive feature because researchers prefer a simpler model which puts more light on the relationship between the output and inputs.

Although the estimator from the lasso appears to be interpretable like the linear regression, the estimator is neither unbiased nor consistent, and it is not possible to interpret it as the OLS estimator. Therefore, while the regression can make a valid inference for the effect of an input, such inference is not possible in the standard penalized regression framework. Recently, some researchers are pursuing valid inference methods for the effect of input in large p setting based on the penalized regression framework (Belloni, Chernozhukov, and Hansen 2011; Belloni and Chernozhukov 2013; Belloni, Chernozhukov, and Hansen 2014; Raskutti, Wainwright, and Yu 2011).

Zou and Hastie (2005) proposed a compromise between the ridge and the lasso, so called *elastic-net*. The elastic-net uses a linear combination of the L_2-penalty and the L_1-penalty as the regularization term:

$$J(\theta) = \alpha\|\theta\|_2^2 + (1-\alpha)\|\theta\|_1,$$

where $\alpha \in [0, 1]$ is an additional hyperparameter which determines the degree of the compromise. They argue that the prediction performance of the elastic-net is expected to be better than that of the lasso if there is a group of variables among which the pairwise correlations are very high.

A *penalized logistic regression* estimator arises as an extension of penalized least squares

methods to the logistic regression framework:

$$\underset{\theta}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \{-\log b_{y_i}(x_i;\theta)\} + \lambda J(\theta) \right],$$

where

$$b_{y_i}(x_i;\theta) = \frac{y_i \exp x_i^T\theta}{1+\exp x_i^T\theta} + \frac{1-y_i}{1+\exp x_i^T\theta}.$$

The regularization term $J(\cdot)$ can be either of the L_2-penalty (Zhu and Hastie 2004), the L_1-penalty (Shevade and Keerthi 2003), or the elastic-net penalty (Waldron et al. 2011).

1.2.6 Tree-based model

Morgan and Sonquist (1963) proposed a simple *tree-based model* that tries to automatically select inputs that are crucial to predict an outcome and flexibly incorporate nonlinearity and interactions of them. The idea is to split the input space into subgroups that can be expressed as a leaf of a decision tree, and then assign a simple predictive value to each subgroup. Subgroups are called *leaves* and *nodes* in the context of the tree-based model.

Classification and regression tree

A popular estimation method for a simple tree-based model called *classification and regression tree* (CART) is introduced by Breiman et al. (1984). Consider the case of regression first. The input space is split into M subgroups $\{R_m\}_{m=1}^M$ and a constant response c_m is assigned to each subgroup as a predictive value for the subgroup inputs:

$$a(X;\theta) = \sum_{m=1}^{M} c_m I(X \in R_m),$$

where $\theta = \left\{\{c_m\}_{m=1}^M, \{R_m\}_{m=1}^M\right\}$, M is a hyperparameter, and each subgroup R_m is defined by a combination of simple decision rules. For example, a subgroup R_m can be defined as a set of inputs that the jth input X_j is over s and the j'th input $X_{j'}$ is over s': $R_m = \{X : X_j > s \quad \text{and} \quad X_{j'} > s'\}$.

The estimation proceeds as follows. Define a pair of half-spaces into which the input space

is divided by a hyperplane $X_j = s$:

$$R_1(j, s) = \{X | X_j \leq s\} \quad \text{and} \quad R_2(j, s) = \{X | X_j > s\}.$$

Then we seek a best splitting hyperplane that minimizes the total loss of the dataset given the constant response rule:

$$\underset{j,s}{\operatorname{argmin}} \left\{ \min_{c_1} \sum_{x_i \in R_1(j,s)} L(y_i, c_1) + \min_{c_2} \sum_{x_i \in R_2(j,s)} L(y_i, c_2) \right\}.$$

A typical loss function is the squared error loss for regression. Having found the hyperplane, we partition the dataset into two regions by the hyperplane and repeat this splitting process on each of the two regions. The process is repeated on all of the resulting regions to grow a tree. A tree is stopped to grow when a designated minimum node size (e.g., ten) is reached. The final tree τ_0 is pruned using a cost-complexity criterion to find an optimal tree.

To state the cost-complexity criterion of the pruning procedure, some notations need to be defined: a subtree τ is defined to be any tree that can be obtained by pruning τ_0; let $|\tau|$ denote the number of leaves in the subtree τ; let $\hat{c}_m$ be the solution that minimizes the loss in the mth node:

$$\underset{c_m}{\operatorname{argmin}} \sum_{x_i \in R_m} L(y_i, c_m).$$

The cost-complexity criterion is defined using these notations with the regularization principle (1.1):

$$C_\lambda(\tau) = \sum_{m=1}^{|\tau|} \sum_{x_i \in R_m} L(y_i, \hat{c}_m) + \lambda |\tau|,$$

where λ is a hyperparameter reflecting the number of subgroups M. Though the loss function in the criterion can be a different loss function from that in the tree growing procedure, usually it is the squared error loss as well. For a given value of λ, one can show that there is a unique subtree τ_λ that minimizes the criterion. In practice, we use the *weakest link pruning* to produce a sequence of subtrees. In the weakest link pruning, we successively collapse the internal node that produces the smallest per-leaf increase in a total loss until a single-node tree is produced.

The subtree that minimizes the criterion in the sequence is known to be the optimal subtree τ_λ.

For K-class classification, the form of candidate functions becomes to

$$b(X;\theta) = \sum_{m=1}^{M} p_m I(X \in R_m),$$

where $\theta = \left\{\{p_m\}_{m=1}^{M}, \{R_m\}_{m=1}^{M}\right\}$ and $p_m = (p_{m0}, p_{m1}, \ldots, p_{mk}, \ldots, p_{mK-1})^T$ is a vector of the conditional probability of the assignment to the K classes in node m. Practically, the estimator of p_{mk}, $\hat{p}_{mk}$, is the proportion of the class k observations in node m. We succesively seek a hyperplane that minimizes the total loss of the dataset given the estimator of p_m, $\hat{p}_m = (\hat{p}_{m0}, \hat{p}_{m1}, \ldots, \hat{p}_{mk}, \ldots, \hat{p}_{mK-1})^T$, in the tree growing procedure. For example, the objective function in the first splitting process is

$$\underset{j,s}{\operatorname{argmin}} \left\{ \sum_{x_i \in R_1(j,s)} L(y_i, \hat{p}_1) + \sum_{x_i \in R_2(j,s)} L(y_i, \hat{p}_2) \right\}.$$

A typical loss function is the log-loss for classification. The pruning procedure is conducted with a loss function that suits classification as well: typically the log-loss again. Consequently, the cost-complexity criterion is slightly changed from that of the procedure for regression,

$$C_\lambda(\tau) = \sum_{m=1}^{|\tau|} \sum_{x_i \in R_m} L(y_i, \hat{p}_m) + \lambda|\tau|,$$

but the remaining part of the procedure is the same.

Random forest

A problem with the simple tree-based model is their high variance of the estimated prediction function. The hierarchical nature of the procedure (i.e., the effect of an error in the top split is propagated down to all of the splits below it) and the lack of smoothness of the prediction surface cause the high variance. *Bootstrap aggregation* or *bagging* (Breiman 1996) is a technique that reduces the variance of an estimated prediction function. Bagging averages predictions that are estimated over a collection of bootstrap samples which are generated from the dataset. If the correlation of pairs of bagged predictions is not perfect, the variance of the bagging estimate is

guaranteed to be smaller than that of the initial estimate. Bagging introduces randomness to the predictions by the use of bootstrap samples and creates a set of predictions that are not perfectly correlated with each other. As bagging works especially well for high-variance and low-bias estimation methods, the power of the tree-based model is highly enhanced by using it.

The larger the correlation of pairs of bagged trees is, the more the benefit of the aggregation is limited. The *random forest* (Breiman 2001) aims to improve the variance reduction property of bagging by lowering the correlation between the trees. The random forest augments randomness of the trees by randomly selecting $\xi \leq p$ of the inputs as candidates for splitting inputs for each split in the tree growing procedure. The number of inputs selected for each split ξ is a hyperparameter for the random forest. Breiman (2001) recommends the default value of $\lfloor p/3 \rfloor$ for regression and $\lfloor \sqrt{p} \rfloor$ for classification for the choice of ξ.

Although the hyperparameter for the tree size was the number of leaves in the CART, the minimum node size or the depth of the tree is commonly used in the random forest. The hyperparameter is usually not tuned via the risk functional estimation to avoid a high computational burden, and it is preset based on existing knowledge. The results are known to be fairly insensitive to particular choices of the hyperparameter (Segal 2004), and Hastie, Tibshirani, and Friedman (2009) evaluates that tuning it does not worth the cost.

Importance sampled learning ensemble

The idea in the *gradient boosting machine* (GBM, Friedman (2001)) and the *stochastic gradient boosting machine* (SGBM, Friedman (2002)) is to *de-correlate* the trees more efficiently than bagging. As these methods are subsumed under the framework of the *importance sampled learning ensemble* (ISLE, Friedman and Popescu (2003)), I introduce the framework here. Regression and classification are explained together.

The procedure of the ISLE is two folds: a preparation of a finite dictionary of trees on which the subsequent aggregation is based using the ISLE generator; an aggregation of the trees in light of the regularization principle. Denote a prediction function of a tree with parameter γ as $\Psi_\gamma^\tau(X)$. The generator is designed to create trees that have a small correlation with each other.

First, let

$$\zeta_0(X) = \begin{cases} \underset{c}{\operatorname{argmin}} \sum_{i=1}^{N} L(y_i, c) & \text{if regression,} \\ \frac{1}{N}\sum_{i=1}^{N}(I(y_i = 0), I(y_i = 1), \ldots, I(y_i = K-1))^T & \text{if } K\text{-class classification.} \end{cases}$$

Then, iterate the next procedure for $d = 1, \ldots, D$. Draw a subsample of $N\eta$ ($\eta \in (0, 1]$) of the dataset, typically with replacement if $\eta = 1$ and without replacement otherwise. Given $\zeta_{d-1}(X)$ and the subsample, we estimate a parameter γ_d of a tree $\Psi^{\tau}_{\gamma_d}(X)$ that predicts an output $Y - \zeta_{d-1}(X)$ from inputs X. The tree is appended to the dictionary as the dth basis function. At the end of the round, update $\zeta_d(X)$ in the following manner:

$$\zeta_d(X) = \zeta_{d-1}(X) + \nu\Psi^{\tau}_{\gamma_d}(X),$$

where ν is a hyperparameter for the *learning rate* of the ISLE generator.

ISLE generator efficiently assembles sufficiently different or de-correlated basis trees via the randomness of subsampling and the learning of $\zeta_d(X)$. The more the subsamples are different from each other, the more the correlation among simple trees is reduced, and the larger the learning rate, the more the procedure avoids a tree similar to those found before.

The aggregation of the trees in the dictionary is accomplished by the regularization technique (1.1):

$$\min_{\{\beta_d\}_{d=0}^{D}} \sum_{i=1}^{N} L[y_i, \beta_0 + \sum_{d=1}^{D} \beta_d \Psi^{\tau}_{\gamma_d}(X)] + \lambda J(\beta).$$

The aggregation technique is called the *post-processing*, and Friedman and Popescu (2003) recommends the use of the L_1-penalty in it. Finally, we get a prediction function $\Psi_\theta(X)$ in the following form:

$$\Psi_\theta(X) = \beta_0 + \sum_{d=1}^{D} \beta_d \Psi^{\tau}_{\gamma_d}(X),$$

where $\theta = \{\{\beta_d\}_{d=0}^{D}, \{\gamma_d\}_{d=1}^{D}\}$.

There are five hyperparameters in the ISLE: a hyperparameter for the tree size, the number of trees to be bagged D, the subsampling ratio for each tree η, the learning rate ν, and the regularization coefficient for the post-processing λ. The tree size is often calibrated by the depth

of the tree. Hastie, Tibshirani, and Friedman (2009) suggests the depth to be six and states that the tuning of the hyperparameter for the depth seldom provides a significant improvement over just using six as the value of it. The number of trees to be bagged D is usually chosen to minimize the estimator of the risk functional.

The basis function generating process of the ISLE is identical to that of the GBM and that of the SGBM if the subsampling ratio η is one and otherwise, respectively. The difference between the ISLE and the boosting machines are in the way that they assemble the basis functions. The ISLE uses the regularization principle more directly than the boosting machines to overcome the disadvantage that they overfit with a growing number of trees to be bagged. Friedman (2002) recommends to set the subsampling ratio to be no more than a half and to be smaller for large sample size: $\eta \leq 1/2$ and $\eta \sim 1/\sqrt{N}$.

The learning rate smaller than one provides regularization through shrinkage, and the recommended value is 0.05 and 0.1 for the GBM and the SGBM, respectively (Friedman 2001, 2002). Lastly, the regularization coefficient for the post-processing λ is chosen to minimize the estimator of the risk functional.

1.2.7 Neural network

The *neural network* was first developed as a model for the human brain (McCulloch and Pitts 1943) and became famous as the basis of deep learning (Hinton, Osindero, and Teh 2006). The modern formulation of a neural network with a *back-propagation algorithm* made the neural network to be a computationally feasible method (Rumelhart, Hinton, and Williams 1986a, 1986b). The neural network attempts to capture nonlinear and interactive effects of inputs on output through the *hidden layer*. As the parameter and the hyperparameter of the model become highly complicated with multiple hidden layers, I concentrate on a single hidden layer neural network here.

A hidden layer consists of derived features or hidden units $Z = (Z_1, Z_2, \ldots, Z_M)$, where M is a hyperparameter of the model. A derived feature is created from a linear combination of the inputs and then the target function $\Psi_\theta(X)$ is modeled as a function of a linear combination of

the derived features: for regression, the model is represented as

$$\begin{cases} a(X;\theta) = \rho(Z^T\beta); \\ Z_m = \sigma(X^T\alpha_m) \text{ for } m = 1,\ldots,M, \end{cases}$$

where $\theta = \{\{\alpha_m\}_{m=1}^M, \beta\}$, $\rho(\cdot)$ is an *output function*, and $\sigma(\cdot)$ is an *activation function*; for K-class classification,

$$\begin{cases} b_k(X;\theta) = \rho_k(V_0, V_1, \ldots, V_{K-1}) \text{ for } k = 0,\ldots,K-1; \\ V_k = Z^T\beta_k \text{ for } k = 0,\ldots,K-1; \\ Z_m = \sigma(X^T\alpha_m) \text{ for } m = 1,\ldots,M, \end{cases}$$

where $\theta = \{\{\alpha_m\}_{m=1}^M, \{\beta_k\}_{k=0}^{K-1}\}$, and $\{\rho_k(\cdot)\}_{k=0}^{K-1}$ are output functions.

The output function is usually the *identity function* for regression, $\rho(Z^T\beta) = Z^T\beta$; and the *softmax function* for classification, $\rho_k(V) = e^{V_k}/\sum_{l=0}^{K-1} e^{V_l}$. The activation function is usually chosen to be the *sigmoid*, $\sigma(X^T\alpha_m) = 1/(1+e^{-X^T\alpha_m})$.

The parameter θ is called *weights* in the context of the neural network, and we seek values for them that make the model fit the dataset well. Generally, neural networks have too many weights and overfit the data at the minimum of the empirical risk functional. Therefore, we apply regularization principle to the weight estimation problem as well. The objective function for the estimation of weights is in the form of the regularization problem (1.1). Typically, the loss function is the squared error loss and the log-loss for regression and classification, respectively, and the regularization is achieved via the L_2-penalty. The regularization technique of using the L_2-penalty in the weight estimation is called *weight decay*.

The number of hidden units M is typically in the range of $[5, 100]$, and the number tends to be large if the number of inputs and the sample size is large. Although one can search for the optimal number to minimize the estimator of the risk functional, it is most common to set a reasonably large number of hidden units and shrink weights with an appropriate regularization technique.

Note that as the neural network model is generally overparameterized, the identification of

the weights is poor. This feature hampers the interpretability of the model and makes the model difficult to apply to biomedical and clinical research except research related to image recognition. Therefore, the neural network is predominantly used for image recognition problems and mostly developing as a method of image recognition (e.g., deep learning models). Using multi-hidden layers with constraints such as *local connectivity* and *weight sharing* on the network, which allow for more complex connectivity but fewer weights, improved the performance of the neural network dramatically in the field of image recognition (LeCun 1989; LeCun et al. 1998). Nevertheless, a multiple hidden layer neural network that suits the situation other than image recognition is not yet sophisticated enough.

Chapter 2

Methods

This chapter is largely based on Hara et al. (2018) except for the section regarding statistical methods.

2.1 Setting

The Japanese government provides a universal health insurance program for all registered inhabitants. In addition, each employer is obliged by law to provide annual health screening to its employees. In Japan, the examination rate of annual health screening provided by a company is very high, 89% on average in 2012, especially with employees' increasing age (Ministry of Health Labour and Welfare 2012). Indeed, in the data I used, more than 80% of employees over 40 years old have undergone the health screening.

2.2 Data

Medical and pharmacy claims data combined with annual health screening results were obtained from Japan Medical Data Center (JMDC). JMDC is a for-profit company that collects data from contracting corporate health insurance programs, which mainly cover employees of large firms. JMDC applies strict policies to protect the privacy of enrollees and medical providers, and all private information that could identify enrollees and medical providers have been removed from the data (Kimura et al. 2010).

Claims data contain information on patients, including gender, birth month, and their diagnostic code, medical institutions, pharmacies, and medical treatments provided. Diagnostic codes and medication codes are classified by the 2003 version of the International Classification of Diseases and Related Health Problems, tenth revision (ICD-10) (WHO 2018a) and the 2016 version of the World Health Organization-anatomical therapeutic chemical (WHO-ATC) code (WHO 2018b), respectively. Enrollees' age was defined as their age in March 2018. Annual health screening results include information on the results of the physical examination and the blood test, whether fasting blood samples were collected, and the answer to a health-related questionnaire including questions for the usage of medications. The study protocol was approved by the Institutional Review Board of the University of Tokyo (application number: 18-40).

2.3 Study population

The baseline study population for condition X (hypertension, diabetes, or dyslipidemia) was defined as beneficiaries (1) who were enrolled in the claims database from 1 April 2016 to 31 March 2018 and whose health screening were sequentially conducted for fiscal year (FY) 2016 and FY2017 (n = 1,040,351), (2) with complete data on self-reported use of blood pressure-lowering drugs, hypoglycemic drugs, and lipid-lowering 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 condition X mentioned later for FY2016 and FY2017 (hypertension, n = 631,289; diabetes, n = 152,368; dyslipidemia, n = 614,434) (Fig. 1).

Employees at high-risk workplace environments (e.g., late-night work, frequent exposure to hazardous substances) are required to undergo a special health screening every six months in addition to the regular annual screening. For enrollees who received screening more than once a year (about 2.9% of observations), I adopted the results of the regular health screening.

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 hospitals (Wilchesky, Tamblyn, and Huang 2004; Bullano et al. 2006; Quan et al. 2009). To make the present study comparable to past research, I restricted the baseline study population to

those who at least once in the FY had visited a clinic/hospital that mainly specializes in internal medicine. Such medical institutions have the function of primary care hospitals in Japan.

2.4 Gold standard

I constructed a gold standard from the results of the annual health screening and discussed algorithms to identify patients' with hypertension, diabetes, and dyslipidemia from medical and pharmacy claims data. I consulted with experts and decided to use two distinct medical examination reports to construct diagnostic criteria for these three conditions. Thus, I used two consecutive FYs (FY2016 and FY2017) of the health screening results to construct the gold standard.

I consulted with experts and defined a gold standard to diagnose each condition in compliance with Japanese guidelines (The Japanese Society of Hypertension (Eds.) 2014; The Japan Diabetes Society (Eds.) 2016; Japan Atherosclerosis Society (Eds.) 2013): for hypertension (1) systolic blood pressure (sBP) $\geq$ 140 mmHg and/or diastolic blood pressure (dBP) $\geq$ 90 mmHg for two straight years, and/or (2) self-report of taking blood pressure-lowering drugs in at least one of the two years; for diabetes, (1) hemoglobin A1c (HbA1c) $\geq$ 6.5% in at least one of the two years and fasting blood glucose (FBG) $\geq$ 126 mg/dL in at least one of the two years, (2) FBG $\geq$ 126 mg/dL for two straight years, and/or (3) self-report of taking hypoglycemic drugs in at least one of the two years; and for dyslipidemia, (1) low-density lipoprotein cholesterol (LDL-C) $\geq$ 140 mg/dL for two straight years, (2) high-density lipoprotein cholesterol (HDL-C) $\leq$ 40 mg/dL for two straight years, (3) triglyceride (TG) $\geq$ 150 mg/dL for two straight years, and/or (4) self-report of taking lipid-lowering drugs in at least one of the two years. Hara et al. (2018) compared the self-report of medication usage with the pharmacy claims-based drug usage and demonstrated that the reliability of the self-report was satisfactorily high.

2.5 Claims-based algorithm

The claims-based algorithm (CBA) was compared with the gold standard. When one thinks of a study focusing on one's target condition, it is often convenient to tabulate the data with an

individual/year in one row and prepare a variable which indicates that the individual has the condition in the year. Thus, for practical purposes, CBAs were based on 1-year medical and pharmacy claims data. I used FY2017 claims data as the source of the CBA and compared it with the diagnosis derived from the gold standard based on health screening results of FY2016-FY2017. Utilization of claims data corresponding to the latter health screening year allows the capture of individuals with new onset of disease. However, considering that the primary focus of this study is chronic diseases, the incidence rate is deemed to be considerably lower than the prevalence, with a subsequent minimal decrease in sensitivity when the former health screening year is used instead. Hara et al. (2018) conducted analyses with changing the year of claims data, which was the basis of the CBAs, and found that the sensitivity decreased with the usage of the data corresponding to the former health screening year as expected. Although there was a small sensitivity improvement by using 2-year claims, use of 2-year claims is a trade-off with the sensitivity increase and the convenience of the variable which indicates the individual's yearly medical condition. Considering these aspects, I concluded that one could practically use 1-year claims data corresponding to the latter health screening year for developing CBAs.

2.5.1 Conventional methods

The CBA research so far has required a knowledge-based condition-specific CBA construction procedure. Typically, researchers selected input variables that are likely to be associated with their target condition and decided how to construct the CBA with the selected variables based on their experience or existing knowledge. They needed to assess a large collection of candidate CBAs to find out a fine-tuned one and iterate the procedure for each target condition. I define conventional methods as methods that select input variables and decide how to incorporate variables into the CBA by hand.

I first developed three case-finding algorithms for each condition. 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); (2) the medication code corresponding to condition X is found in the claims at least once (medication code-based CBA), and (3) the diagnostic

code and the medication code corresponding to condition X are both found in the claims data at least once (combined CBA). The diagnostic codes corresponding to hypertension, diabetes, and dyslipidemia were, respectively, defined as ICD-10 codes I10-I15, E10-E14, and E78. The medication codes corresponding to hypertension, diabetes, and dyslipidemia were, respectively, defined as WHO-ATC codes C08 (calcium channel blockers) and C09 (agents acting on the renin-angiotensin system), A10 (drugs used in diabetes), and C10 (lipid-modifying agents). I designate these algorithms as baseline CBAs in the conventional methods.

Next, I examined the effects of slightly modifying the baseline selection rule. I changed the threshold for each algorithm by requiring the presence of two or three diagnostic and/or medication codes to consider a patient positive. I counted observations of diagnostic or medication codes on claims as one occurrence when information was accrued from the same month. Thus, the presence of two or three codes indicates the existence of the codes in two or three distinct months. In addition, I broadened the definition of medication codes for hypertension to include C02 (antihypertensive drugs), C03 (diuretic drugs), and C07 (beta-blocking agents) as some physicians may prescribe these drugs for blood-pressure lowering as well.

2.5.2 Statistical methods

Statistical methods such as regression and statistical learning methods can foster the development of CBAs. However, regression and some statistical learning methods are poor at dealing with sparse high-dimensional input variables. Consequently, they require a variable selection before the implementation. I applied (1) regression model, (2) discriminant analysis, and (3) generalized additive model (GAM) to a dataset that input variables were selected according to each condition.

To bypass a somewhat cumbersome task of selecting variables that are likely to be associated with each target condition and constructing a satisfactory CBA from the selected variables, I devised methods by which a CBA is fine-tuned regardless of the level of knowledge and without modification of the CBA construction procedure across different conditions.

I applied (1) logistic regression, (2) k-nearest neighbor (kNN), (3) support vector machine (SVM), (4) penalized regression, (5) tree-based model, and (6) neural network to a dataset that input variables were chosen to be common to all target conditions. Although regression methods

can be used when the number of the input variables is smaller than the sample size and the input variables with perfect colinearity were trimmed in advance, their predictive property is expected to be poor. To examine this point, I included a logistic regression to the models. The statistical learning methods elected are capable of handling the sparse high-dimensional input variables.

2.6 Statistical analysis

I quantified the goodness of CBAs by association measures, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), receiver operating characteristic (ROC) curve, and area under the ROC curve (AUC). I defined sensitivity, specificity, PPV, and NPV, as follows: sensitivity was defined as the proportion of enrollees who were identified as having a disease by the CBA among those who were assessed as having that disease by the gold standard; specificity was defined as the proportion of enrollees who were identified as not having a disease by the CBA among those who were assessed as not having that disease by the gold standard; PPV was defined as the proportion of enrollees who were assessed as having a disease by the gold standard among those who were identified as having that disease by the CBA; and NPV was defined as the proportion of enrollees who were assessed as not having a disease by the gold standard among those who were identified as not having that disease by the CBA.

Before the analysis, the dataset was randomly divided into three parts: a test set (25%), a training set (50%), and a validation set (25%). Association measures of the CBA were assessed using the test set. Note that the test set was never used for the parameter or the hyperparameter estimation.

As the computational burden of some statistical methods without a condition-specific variable selection was prohibitive for large sample size, I randomly drew 25% of the enrollees for the analysis of hypertension and dyslipidemia except for conventional methods. Although the analysis with a condition-specific variable selection can be accomplished by the full sample size, it is also conducted using the same 25% random sample to make the results comparable to the analysis without it.

All statistical analysis was conducted using R version 3.5.1 (R Core Team 2018). R code will be available at https://github.com/harakonan/research-public/tree/master/cba after the pub-

lication of the study.

2.6.1 Conventional methods

The sensitivity, specificity, PPV, and NPV were estimated for the CBA derived from the conventional methods, and 95% confidence intervals (CIs) for them were calculated using exact binomial confidence limits (Collet 1999). I calculated the association measures and 95% CIs for the association measures using the *epiR* package (Stevenson et al. 2018).

Sensitivity analysis for conventional methods

In order to evaluate to what extent baseline CBAs in conventional methods is applicable to a wide range of populations, I implemented a multipronged strategy for sensitivity analysis with respect to the underlying study population. At first, the following study populations were considered: (1) enrollees who had visited any clinic/hospital at least once in FY2017; and (2) all enrollees including those who had not visited any clinic/hospital in FY2017. An individual's health condition cannot be identified from the claims data unless they have visited a clinic/hospital. Moreover, those who had not visited a clinic/hospital which is primarily providing internal medicine seem to have a low possibility of being diagnosed with hypertension, diabetes, and/or dyslipidemia. Consequently, I predict that the sensitivity will decrease by this study population expansion. Next, I examined the change in association measures by focusing on a high-prevalence population: enrollees aged 50 years or older in the baseline study population.

Additionally, to take into account the possibility of underdiagnoses of our elected gold standard, sensitivity analysis regarding the gold standard was conducted as well. As hypertension, diabetes, and dyslipidemia can be diagnosed by a single physical or laboratory examination, there is a possibility that enrollees were underdiagnosed with the gold standard. I relaxed the criteria of the gold standard accordingly: the criterion of the physical examination part of hypertension was modified to sBP $\geq$ 140 mmHg or dBP $\geq$ 90 mmHg in at least one of the two years; the criterion of the laboratory examination part of diabetes was modified to HbA1c $\geq$ 6.5% in at least one of the two years; and the criterion of the laboratory examination part of dyslipidemia was modified to LDL-C $\geq$ 140 mg/dL, HDL-C $\leq$ 40 mg/dL, or TG $\geq$ 150 mg/dL in at least one

of the two years. A more extensive collection of sensitivity analysis than this dissertation was analyzed and discussed elsewhere (e.g., the paper used claims-based medication data instead of self-reports in the gold standard) (Hara et al. 2018).

2.6.2 Statistical methods

All association measures were calculated for the CBA derived from statistical methods. A prediction function needs to be derived from the statistical model to calculate association measures. As the current problem is a two-class classification problem, I estimated a prediction function that outputs the score of the propensity for having a disease given a set of input variables. The outcome variable in hand is a binary indicator of having a disease that is assessed by the gold standard.

If the model involves a hyperparameter to be tuned, the training set and the validation set were used for the tuning. For each candidate value of the hyperparameter, an estimation of the parameter of the model is conducted with the training set. Given the estimated parameter, the AUC of the model is computed using the validation set. Then, the hyperparameter is chosen to be the value that maximizes the AUC. If computationally feasible, tenfold cross-validation with a combined set of training and validation set (simply "combined training set" in what follows) was used to estimate the expected value of the AUC. After the hyperparameter determination, the combined training set was used to estimate parameters for the prediction function. When no hyperparameter tuning is required, the combined training set was used to estimate parameters in the prediction function from the beginning.

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 is moved over the range of all possible scores. The AUC was calculated from the resulting ROC curve, and 95% CI for the AUC was calculated with Delong's method (DeLong, DeLong, and Clarke-Pearson 1988).

In the end, a representative point of sensitivity and specificity on the ROC curve is chosen based on the Youden index (Youden 1950). The Youden index reflects the intention of minimizing misclassification rates and is considered to be more clinically meaningful compared to the other

common method in which the point on the ROC curve closest to the upper left corner is chosen (Perkins and Schisterman 2006). The PPV and NPV were calculated according to the representative point. I calculated 95% CIs for the sensitivity, specificity, PPV, and NPV were calculated with 200 bootstrap resampling and the averaging methods described by Fawcett (2006). I drew the ROC curve and calculated the association measures and 95% CIs for the association measures using the *pROC* package (Robin et al. 2011).

I did not take into account the uncertainty of the initial prediction function estimation procedure in the 95% CIs for the association measures. As the final product of this study is a particular algorithm to find out individuals with a target disease, the algorithm or the prediction function estimated in this study is assumed to be as it is, and the uncertainty of the prediction function estimation procedure is not a concern. If a future study begins from the prediction function estimation instead of just using the prediction function provided in this study, the 95% CIs may be biased.

Statistical methods with a condition-specific variable selection

I selected two set of input variables for the regression model: one that mimics the input variables used in the conventional methods, the number of observations of each of the diagnostic code and the medication code (model 1); another that examines the effect of incorporating demographics, age and gender besides the input variables selected in model 1 (model 2).

Note that the maximum possible number of the variable regarding the diagnostic/medication code is twelve as I derived CBAs based on 1-year medical and pharmacy claims data. I refer to the dataset used for model 2 as a "condition-specific dataset" in what follows. The condition-specific dataset was used for the discriminant analysis and the GAM.

Regression model

The outcome variable was regressed on the selected input variables to generate a prediction function. Linear and logistic regressions were performed for model 1 and model 2. The analysis of the logistic regression was implemented by the *mnlogit* package (Hasan, Wang, and Mahani 2016).

Discriminant analysis

Three types of discriminant analysis were conducted: linear discriminant analysis (LDA, Fisher (1936)); flexible discriminant analysis (FDA, Hastie, Tibshirani, and Buja (1994)); and penalized discriminant analysis (PDA, Hastie, Buja, and Tibshirani (1995)). Multivariate adaptive regression splines of Friedman (1991) with the second degree of interaction were used as the basis function in the FDA. In the PDA, a linear basis with the L_2-penalty was used, and the regularization coefficient was determined by cross-validation. Posterior class probabilities based on a Gaussian assumption described by Hastie, Tibshirani, and Buja (1994) were used to generate a prediction function. The discriminant analysis was conducted by the *mda* package (Hastie et al. 2017).

Generalized additive model

The additive logistic regression model with a cubic smoothing spline for each input variable except gender was used in the generalized additive model (GAM) (Hastie and Tibshirani 1986, 1990). The hyperparameter, the degrees of freedom for each smoothing spline, was set to the same value for all splines. I selected three values for the degrees of freedom, four, six, and eight, to generate three GAM prediction functions. The analysis of the GAM was implemented by the *gam* package (Hastie 2018).

Statistical methods without a condition-specific variable selection

A dataset that consists of age, gender, and all ICD-10/WHO-ATC codes with a letter followed by two digits as input variables was set up. This dataset can be constructed without prior knowledge about the relationship between diagnostic/medication codes and the target condition and can be used regardless of the target condition. I refer to the dataset as a "general dataset" in the following.

Logistic regression

A logistic regression model was fitted to the dataset that was trimmed properly. The trimming was applied to the combined training set of the general dataset as it was used for the parameter

estimation of the logistic regression model.

At first, input variables that provide no information, i.e., the value of the input variable was the same for all observations, were removed. Then, the input variables that can be written as a linear combination of other input variables were removed. Theoretically, the dataset after these two trimming practices can be used to generate a prediction function. Nevertheless, the convergence of the estimation procedure can be very slow if the correlation between an input variable and some linear combination of other input variables is too high. Hence, I also removed the input variables that were highly correlated with some linear combination of other input variables to sidestep an excessively long computational time. Although this additional trimming practice is frequently used to deal with the model that is difficult to estimate, the practice can render a better predictive property to the model through a variance reduction of the estimator of the remaining parameters. Consequently, the resulting association measures can be overestimated. The following statistical learning methods internally achieve a sort of dimension reduction in the input space and proficiently get around these issues.

k-nearest neighbor

The Euclidean distance with raw or standardized (i.e., re-scaled to have mean zero and variance one) input variables was adopted as a distance metric for the 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, Fix and Hodges (1951)) and (2) the inverse distance weighted frequency of the class of the k-nearest neighbors (IDW, Shepard (1968)) composed a prediction function. The analysis of the kNN was implemented by the *fastknn* package (Pinto 2018).

Support vector machine

A linear basis function with a hinge or squared hinge loss was adopted in the SVM (Cristianini and Shawe-Taylor 2000). 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 (Helleputte 2017).

Penalized regression

From the penalized regression, logistic regressions with the L_2-penalty (logistic ridge, Zhu and Hastie (2004)), the L_1-penalty (logistic lasso, Shevade and Keerthi (2003)), and the elastic-net penalty (logistic elastic-net, Waldron et al. (2011)) were applied. The regularization coefficient and the elastic-net mixing parameter were determined by cross-validation. The analysis of the penalized regression model was implemented by the *glmnet* package (Friedman, Hastie, and Tibshirani 2010).

Tree-based model

Two types of tree-based models were applied: the random forest (Breiman 2001) and the importance sampled learning ensemble (ISLE, Friedman and Popescu (2003)). The minimum node size was set to ten for each tree, and two-hundred trees were bagged in the random forest. The number of variables selected for each split was first set to the value recommended by Breiman (2001), $\lfloor \sqrt{p} \rfloor$, where p is the number of input variables, and then tuned using the validation set. I grew a probability forest as in Malley et al. (2012) to generate a prediction function. The analysis of the random forest was implemented by the *ranger* package (Wright and Ziegler 2017).

There are five hyperparameters in the ISLE: a hyperparameter for the tree size, the subsampling ratio for each tree, the learning rate, the number of trees to be bagged, and the regularization coefficient for the post-processing. I adopted the depth of the tree as the hyperparameter for the tree size and fixed it to be six as is recommended by Hastie, Tibshirani, and Friedman (2009).

The ISLE can be primarily divided into two according to the subsampling ratio for each tree: the basis function generating process of the ISLE is identical to that of the gradient boosting machine (GBM, Friedman (2001)) and that of the stochastic gradient boosting machine (SGBM, Friedman (2002)) if the subsampling ratio is one and otherwise, respectively. Denote the ISLE with the subsampling ratio equals to one and less than one by ISLE-GBM and ISLE-SGBM, respectively.

For the ISLE-GBM, I selected the learning rate to be 1 (i.e., no shrinkage in the learning process of the ISLE generator) and 0.05 (the value recommended for the GBM by Friedman

(2001)). For the ISLE-SGBM, I fixed the learning rate to be 0.1, the value recommended for the SGBM by Friedman (2002), and selected the subsampling ratio to be 0.5 and 0.1.

The remaining two hyperparameters, the number of trees to be bagged and the 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, and the value with the maximum cross-validation AUC was chosen as the regularization coefficient for the prediction function. The L_1-penalty was adopted in the post-processing as is recommended by Friedman and Popescu (2003). The analysis of the ISLE was implemented by the *xgboost* package (Chen and Guestrin 2016).

Neural network

A single hidden layer neural network was applied with a different number of hidden units: five, ten, and twenty. All hidden units were fully connected with the nodes in the input and output layer. 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 (Venables and Ripley 2002).

Chapter 3

Results

3.1 Summary statistics

Table 1 tabulates summary statistics of 944,717 enrollees' characteristics and health screening 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. The proportion of enrollees visiting clinics/hospitals during the year was 85% for any clinics/hospitals and 67% for the primary care clinics/hospitals. Blood pressure, FBG, HbA1c, LDL-C, and TG values have increased from FY2016 to FY2017.

To summarize the input variables regarding diagnostic/medication codes in the condition-specific dataset, I drew a probability mass function of the number of observations of diagnostic/medication code corresponding to hypertension, diabetes, and dyslipidemia for the baseline study population (Figure 2). The proportion of enrollees whose claims contain a diagnostic/medication code corresponding to hypertension, diabetes, and dyslipidemia was 24%/21%, 14%/7%, and 25%/15%, respectively.

Next, I summarized the input variables regarding diagnostic/medication codes in the general dataset. 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 baseline study population. Cumulative distribution of the computed proportion was tabulated separately for ICD-10 codes and WHO-ATC codes (Table 2). 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.

The numbers of ICD-10 codes and WHO-ATC codes appeared in the general dataset for the baseline study population were 1333 and 92, respectively. Nearly 90% of ICD-10 codes appeared in the dataset were only observed for less than 1% of enrollees, and more than half of WHO-ATC codes appeared in the dataset were observed for less than 5% of enrollees. As a whole, the input variables in the general dataset were sparse and high-dimensional: a small fraction of the entries of the input variables was non-zero, and the number of the input variables were over 1400.

3.2 Conventional methods

Table 3A to 3C report the association measures and their 95% CIs for the CBAs derived from the conventional methods according to hypertension (Table 3A), diabetes (Table 3B), and dyslipidemia (Table 3C). As the test set (25% of the baseline study population of each condition) was used in the calculation of the association measures, the sample size was 157,822, 38,092, and 153,608 for hypertension, diabetes, and dyslipidemia, respectively. For them, the prevalence which was determined by the gold standard for each condition was 25.4%, 8.3%, and 38.7% for hypertension, diabetes, and dyslipidemia, respectively.

In the baseline diagnostic code-based (combined) CBA, the sensitivity, the specificity, PPV, and NPV were 80.4%, 95.1%, 84.9%, and 93.4% (74.4%, 98.1%, 93.1%, and 91.8%) for hypertension, 91.1%, 92.8%, 53.4%, and 99.1% (79.2%, 99.6%, 94.7%, and 98.2%) for diabetes, and 49.2%, 90.1%, 75.8%, and 73.7% (35.8%, 97.0%, 88.2%, and 70.5%) for dyslipidemia. When increasing the threshold of the number of observations of diagnostic and/or medication codes required to consider a patient positive, the sensitivity and NPV decreased, whereas the specificity and PPV increased. The direction of the change of the association measures was the same if alternative medication codes were used for hypertension.

Sensitivity analysis for conventional methods

Table 4A to 4C present the sample size, the prevalence, the association measures, and their 95% CIs of each sensitivity analysis for the baseline CBA according to hypertension (Table 4A), diabetes (Table 4B), and dyslipidemia (Table 4C). The sensitivity decreased when the study population was expanded to include all people (hypertension, 66%-71%; diabetes, 74%-85%; dyslipidemia, 28%-38%) and increased when the study population was constrained to enrollees aged 50 years or older (79%-84%; 80%-92%; 45%-58%). The sensitivity decreased when the criteria in the gold standard for each condition were relaxed (58%-64%; 73%-87%; 26%-38%).

3.3 Statistical methods

Statistical methods with a condition-specific variable selection

Table 5A to 5C show the association measures and their 95% CIs for the CBAs derived from the statistical methods with a condition-specific variable selection according to hypertension (Table 5A), diabetes (Table 5B), and dyslipidemia (Table 5C). Figure 3A to 3C display the ROC curve for the corresponding prediction functions (hypertension, Figure 3A; diabetes, Figure 3B; dyslipidemia, Figure 3C).

The AUC of the regression model increased by adding the demographics to the input variables: hypertension, the AUC of model 1 (model 2), .895-.897 (.924-.925); diabetes, .946-.947 (.958-.962);dyslipidemia, .709-.710 (.738-.739). The AUC of the discriminant analysis was .928-.929 for hypertension, .963 for diabetes, and .758 for dyslipidemia. The AUC of the GAM was .925 for hypertension, .962 for diabetes, and .739-.746 for dyslipidemia.

The model with the highest AUC, the GAM with eight (six) degrees of freedom for hypertension and diabetes (dyslipidemia), achieved the following association measures at the representative coordinate on the ROC curve: hypertension, sensitivity 81.4%, specificity 95.1%, PPV 85.0%, NPV 93.7%; diabetes, 90.8%, 93.5%, 55.7%, 99.1%; dyslipidemia, 49.6%, 90.6%, 77.0%, 73.9%.

Statistical methods without a condition-specific variable selection

Table 6A to 6C show the association measures and their 95% CIs for the CBAs derived from the statistical methods without a condition-specific variable selection according to hypertension (Table 6A), diabetes (Table 6B), and dyslipidemia (Table 6C). Figure 4A to 4C display the ROC curve for the corresponding prediction functions (hypertension, Figure 4A; diabetes, Figure 4B; dyslipidemia, Figure 4C).

The AUC of the logistic regression was .915 for hypertension, .936 for diabetes, and .743 for dyslipidemia. The AUC of the kNN with raw (standardized) input variables was .914-.915 (.855-.856) for hypertension, .942 (.888-.889) for diabetes, and .739 (.677-.680) for dyslipidemia. The AUC of the SVM was .914-.919 for hypertension, .944-.950 for diabetes, and .724-.749 for dyslipidemia.

In the penalized regression, the AUC of the logistic ridge (the logistic lasso and the logistic elastic-net) was .893 (.923-.924) for hypertension, .930 (.961) for diabetes, and .725 (.748-.753) for dyslipidemia. In the tree-based model, the AUC of the random forest (the ISLE) was .923 (.928-.930) for hypertension, .958-.960 (.963-.965) for diabetes, and .760-.761 (.767-.772) for dyslipidemia. The AUC of the neural network was .910-.914 for hypertension, .919-.939 for diabetes, and .739-.745 for dyslipidemia.

The model with the highest AUC, the ISLE-GBM with a shrinkage in the learning process for all conditions, achieved the following association measures at the representative coordinate on the ROC curve: hypertension, sensitivity 81.5%, specificity 95.0%, PPV 85.0%, NPV 93.7%; diabetes, 90.3%, 94.1%, 58.2%, 99.1%; dyslipidemia, 50.1%, 90.6%, 77.2%, 74.1%.

Chapter 4

Disscusion

Using health screening results as the source of the gold standard, I focused on the CBA for identifying patients with three common chronic medical conditions, hypertension, diabetes, and dyslipidemia, and demonstrated (1) the association measures of the CBAs derived from the conventional methods across a large and wide range of populations, and (2) the association measures of the CBAs derived from statistical methods with and without a condition-specific variable selection.

4.1 Conventional methods

I begin a discussion from the baseline CBAs in the conventional methods as a benchmark. For hypertension, all association measures were already acceptably high in diagnostic code-based CBA, and the specificity and PPV were boosted while maintaining the sensitivity around 75% in combined CBA. For diabetes, combined CBA discriminated diabetic patients from nondiabetic individuals accurately while diagnostic code-based CBA fell short of a satisfactory level of the PPV. In contrast to the CBAs for hypertension and diabetes, I could not achieve a satisfactorily high level of the sensitivity for dyslipidemia in any baseline CBA.

For hypertension and diabetes, the association measures obtained in this research were overall higher than those obtained in North American studies (Rector et al. 2004; Robinson et al. 1997; Wilchesky, Tamblyn, and Huang 2004; Bullano et al. 2006; Quan et al. 2009; Tessier-Sherman et al. 2013). The studies all included sensitivity and specificity of hypertension, and no algorithm

reached the sensitivity and specificity over 80% and 95% simultaneously. Likewise, no algorithm reached the sensitivity around 80% and specificity above 95% simultaneously among the three studies focusing on diabetes (Rector et al. 2004; Robinson et al. 1997; Wilchesky, Tamblyn, and Huang 2004). The fact indicates that research using Japanese claims data can find out the patients with hypertension and those with diabetes accurately, and the findings from the research are credible.

The direction of the change of the association measures when moving from diagnostic code-based CBA to combined CBA and when increasing the threshold in the CBA was the same: the sensitivity and NPV decrease, and the specificity and PPV increase with the change. However, the magnitude of the change of the threshold was small compared to the change of the code to use in the CBA. The increase of the threshold triggered a sizable sensitivity decrease and a specificity increase in the past study which investigated the effect of the change of the threshold (Bullano et al. 2006). The relatively small change in this research is possible because, in Japan, diagnostic codes tend to be maintained in the system once they are registered in the claims system. Additionally, the high frequency of medical institution visits in Japan may further attenuate the impact of the threshold change.

As I expanded the study population to include all enrollees from the baseline study population, the sensitivity decreased. Because enrollees who had not visited any clinic/hospital in FY2017 yielded no claims in the year, there is an increase in the number of enrollees whose corresponding condition cannot be assessed using CBAs by construction. The decrease of the sensitivity was mild for hypertension (74%-80% to 66%-71%) and diabetes (79%-91% to 74%-85%), while the decrease was sizable for dyslipidemia despite the low starting point (36%-49% to 28%-38%). I thus speculate that most of those with hypertension or diabetes are visiting medical institutions, whereas only a fraction of those with dyslipidemia is visiting medical institutions. This may be because the consequence of dyslipidemia is less noticeable than those of hypertension and diabetes in Japan.

Based on the considerations so far, I make a proposal for the suitable CBA among conventional methods for each condition. There are largely three types of studies: (1) studies that lay weight on the sensitivity and PPV, e.g., studies for estimating prevalence; (2) studies that require

both high PPV and high NPV, e.g., studies that compare people with and without the target condition; and (3) studies that attach importance only on the PPV ,e.g., studies that require an accurate identification of patients with the target condition. In the case of hypertension, it would be better to use diagnostic code-based CBA in the first category and combined CBA in the last category. There is a trade-off between the PPV and NPV in selecting diagnostic code-based CBA or combined CBA in the second category. Next, in the case of diabetes, I recommend the use of combined CBA in any study category as the PPV of diagnostic code-based CBA is unsatisfactory low. Lastly, in the case of dyslipidemia, because the sensitivity and NPV is unsatisfactory low for any CBA, the use of any CBA in the first and second categories cannot be justified. Thus, I recommend the use of combined CBA to obtain high PPV and conduct research in the last category. Besides, in all conditions, one can obtain even higher PPVs by increasing the threshold of the number of observations of diagnostic and medication codes used to consider a patient positive in the CBA.

4.2 Statistical methods

Statistical methods with a condition-specific variable selection

The AUC of the regression model with the diagnostic/medication code was augmented by adding age and gender to the input variables. Age and gender are known to affect association measures of CBAs for various conditions (Muhajarine et al. 1997; Freeman et al. 2000; Taylor, Fillenbaum, and Ezell 2002; Nattinger et al. 2004; Gold and Do 2007; Østbye et al. 2008; Quan et al. 2009; Kawasumi et al. 2011; Walraven and Colman 2016). As the sensitivity analysis for a high age population implicates, age seems to be an important factor by which the association measures are affected in this study as well. Thus, the rise of the AUC can be understood reasonably.

The consideration of nonlinearity and interactions within the selected input variables using the discriminant analysis and the GAM improved the AUC for dyslipidemia but not for hypertension and diabetes. Increasing the complexity of the model did not contribute to an additional performance gain except for the FDA for dyslipidemia. It seems to be difficult to boost an already high level of the AUC by devising the nonlinearity and interactions of the input variables in the

model.

Statistical methods without a condition-specific variable selection

The use of the general dataset rather than the condition-specific dataset degraded the AUC of the logistic regression but for dyslipidemia. The inconsistency of the trend of the AUC demonstrates the trade-off between the accuracy and the variance of the prediction function. When the number of the input variables of the prediction model becomes large, 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. If the factors of being diagnosed as the target condition are successfully captured in the condition-specific dataset (i.e., a high AUC is achieved by the condition-specific dataset), a regression method with the general dataset suffers from an accuracy deterioration because the effect of the variance increase dominates that of the accuracy gain. Conversely, when the factors of being diagnosed as the condition are not sufficiently covered by the condition-specific dataset, the effect of the accuracy gain outweighs that of the variance increase. Likewise, the AUC for hypertension and diabetes and the AUC for dyslipidemia had different trends for the statistical learning methods with the general dataset reflecting the difference of the difficulty of catching the factors of being diagnosed as the condition.

The AUC of the kNN with raw input variables was as good as that of the logistic regression with the general dataset, but that of the kNN with standardized input variables was lower than it. 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 mean or low standard deviation such as age and gender than otherwise. Although the kNN had established an era in the field of image recognition by the invention of the tangent distance (Simard, LeCun, and Denker 1992), 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, e.g., principal component analysis (Mardia, Kent, and Bibby 1979), before measuring the distance. Though I do not probe further in this study, this is one direction of

future research.

Among the logistic regression, the SVM, and the penalized regression, the logistic lasso and the logistic elastic-net achieved superior AUC to the others. In the remaining, the logistic regression tended to rank first, the SVM second, and the logistic ridge third. They are all linear in parameters model with different loss and penalty functions. The logistic regression and the penalized regression use the log-loss, while the SVM uses the (squared) hinge loss. Four different penalty functions are used: the logistic regression, zero penalties; the SVM and the logistic ridge, the L_2-penalty; the logistic lasso, the L_1-penalty; the logistic elastic-net, the elastic-net penalty.

The methods using the L_1-penalty is better suited to sparse and high-dimensional situations than the methods using zero penalties or the L_2-penalty because of the selection of the effective 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 (Donoho and Elad 2003; Donoho 2006; Candes and Tao 2007). 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 (Zou and Hastie 2005), and usually the diagnostic codes and the medication codes corresponding to the target disease are highly correlated, I could not boost the AUC by the elastic-net compared to the lasso.

While the hinge losses give zero penalties to points correctly classified and outside the margin, the log-loss gives continuously decreasing penalty 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.

The use of the hinge losses with the L_1-penalty may further boost the performance in my setting. However, as the hinge losses and the L_1-penalty are computationally much harder to deal with than the log-loss and the L_2-penalty, respectively, handling both of them simultaneously with high-dimensional data is very difficult especially for large sample size. Devising a method

that can overcome this computational obstacle is another direction of future research.

The tree-based model and the 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 and was at least as good as the logistic regression with the condition-specific variable selection. Among the tree-based model, 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 (Friedman and Popescu 2003; Hastie, Tibshirani, and Friedman 2009). Therefore, two components of the ISLE are contributing to the superior performance of it to that of the random forest: the learning term in the basis function generating process; the 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 with the general dataset. The performance of the neural network was much lower in the preliminary investigation that used smaller sample size (e.g., the sample size of 10,000 and 50,000 for each condition before dividing the dataset into three parts). The number of parameters in the neural network is nearly 7500, 15,000, and 30,000 for five, ten, and twenty hidden units, respectively. Though the use of weight decay should alleviate the overfitting of the parameters to some extent, the sample size of 150,000 may be still 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 (LeCun 1989; LeCun et al. 1998), 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.

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 efficient CBA construction

procedure, I briefly introduce the concepts that are expected to become important in the future accuracy pursuit of CBAs. The first one is more complicated and sophisticated learning models flourished in the field of machine learning including deep learning models (Hinton, Osindero, and Teh 2006). Secondly, pre-processing techniques that transform datasets *ex-ante* to utilize the power of learning machines more efficiently. There are largely two approaches for pre-processing: methods that deal with imbalanced datasets (Chawla et al. 2002) and methods that perform variable and feature selection (Guyon and Elisseeff 2003). Lastly, error analysis in the performance analysis and debugging step of model building (Amershi et al. 2015). How one can successfully use these methods in CBA or, more broadly, claims data situation should be a worthwhile subject to be pursued.

In sum, the penalized regressions other than ridge and the tree-based models, which are the leading statistical learning methods, achieved AUCs comparable to the logistic regression with a knowledge-based condition-specific variable selection, and the level of the AUC was satisfactory for hypertension and diabetes.

4.3 An efficient CBA study

From the considerations so far, I propose a two-step course of action for an efficient CBA research. The first step is to prepare an efficient gold standard construction environment to sidestep chart reviewing. This can be achieved by the use of regularly collected data like annual health screening results, which are used in this study. EHRs and disease registries are possible candidates along this line. For example, an increasing number of phenotyping algorithms may well function as gold standards for CBA research when EHRs are available. Besides, cancer registries can be used to conduct comprehensive CBA research for various cancers. In fact, some CBA research is using health screening results (Tessier-Sherman et al. 2013; Hara et al. 2018), blood test results from EHRs (Gorina and Kramarow 2011; Yamana et al. 2016), and disease registries (Freeman et al. 2000; Taylor, Fillenbaum, and Ezell 2002; Nattinger et al. 2004; Gold and Do 2007; Kawasumi et al. 2011; Chan et al. 2016).

The second step is to use a condition-invariant procedure in the CBA construction. From this study, I recommend using the penalized regressions other than ridge or the tree-based models

with input variables as age, gender, and all ICD-10/WHO-ATC codes with a letter followed by two digits to generate a prediction function that outputs the score of the propensity for having a disease. This procedure is expected to yield an AUC that is comparable to the AUC of the logistic regression with a knowledge-based condition-specific variable selection. Although the suitable statistical learning method may change depending on the selected input variables, one can also include additional enrollee characteristics, ICD-10/WHO-ATC codes with three or more digits, and procedure codes to enhance one's AUC further.

Once a broad set of input variables are selected, researchers can uniformly apply the procedure to construct a prediction function for each of their target conditions and compare it against their gold standard that is constructed from the regularly collected data. All coordinates on the ROC curve can be realized by the CBA induced by the prediction function. The course of action should considerably encourage the implementation of CBA research.

4.4 Strength and weakness

The use of regularly collected data such as the routine health screening results as the source of the gold standard is a novel approach in the literature of CBA. There are advantages of adopting health screening results over the standard of chart review. First, once the gold standard for the target condition is defined, one can systematically acquire the gold standard diagnosis of enrollees without relying on chart reviewers' decision on diagnosis. Second, it takes much less time to run a computer program on health screening results than review charts to obtain the gold standard diagnosis. Third, while the chart review disregards the relevant information which is included in the charts of other medical institutions that is not on the review list, health screening captures the required information for the present three conditions.

The use of statistical learning methods in the CBA construction procedure is an innovative strategy in the literature. Researchers needed to select input variables and decide how to incorporate variables in the CBA with existing knowledge on a case-by-case basis. They may not be so confident about whether the resulting CBA is sufficiently capturing features of the target condition, especially if they failed to attain a satisfactory performance by the CBA. Consequently, it is necessary to conduct a tedious comparison of a large collection of

knowledge-based candidate CBAs to alleviate the uneasiness. An appropriate statistical learning method overcomes these issues proficiently: researchers only need to select variables that can be uniformly applied to all conditions and the variables that are crucially related to the target condition will be incorporated in the model automatically.

Several caveats are in order. At first, the diagnostic accuracy of the gold standard is a concern. Hara et al. (2018) conducted sensitivity analysis to allow for the possibility of underdiagnosis of the target condition resulting from the elected gold standard and quantified the extent to which the association measures are robust to the ways of constructing the gold standard. Besides, the selection of health screening results over medical charts do not necessarily compromise the accuracy of the diagnosis as the enrollees' physical examination results, their blood test results, and their medication use are a key to diagnosing the patient with the present three conditions. For diabetes, this conjecture is backed by the factors included in the phenotyping algorithm for the identification of patients with diabetes (Upadhyaya et al. 2017). One can think of obtaining a more accurate diagnosis with a cohort in which participants are screened and confirmed their medical conditions periodically. Though this may be an ideal way of dealing with the gold standard problem, launching such a cohort of an adequate size from the very beginning is unrealistic for most of the researchers as it demands an innumerable amount of resources.

Second, there is a two-dimensional generalizability issue: the study population only covers regular employees; the research only dealt with three conditions, hypertension, diabetes, and dyslipidemia. Besides the problem on the generalizability of the value of association measures computed in this research, researchers need to think of the following three questions when they try to use the series of techniques evaluated here in a different setting: (1) Is it possible to construct a gold standard efficiently?; (2) Are input variables selected for the condition-invariant procedure sufficient?; (3) What kind of learning method should be used? Although I already discussed the elements to consider in answering these questions, I 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.

Chapter 5

Conclusion

The dissertation showed that one can (1) construct fine-tuned CBAs using a statistical learning method without knowledge for target conditions and condition-specific modifications of the CBA construction procedure and (2) make an assessment of the usability of CBAs in a large population efficiently when regularly collected data as a source of the gold standard is available. I believe that the series of techniques evaluated in the study should become essential in future CBA research.

Acknowledgements

I greatly appreciate my supervisor, Professor Yasuki Kobayashi, for his guidance and persistent support throughout the doctoral course. I also appreciate my colleagues Yuki Ito and Atsushi Miyawaki for their valuable comments on statistical learning methods. Chapter 2 is largely based on Hara et al. (2018) except for the section regarding statistical methods. I thank the co-authors of the article (Jun Tomio, Thomas Svensson, Rika Ohkuma, Akiko Kishi Svensson, and Tsutomu Yamazaki) for their permission to use it in the dissertation.

Bibliography

Abaluck, Jason, and Jonathan Gruber. 2016. "Evolving choice inconsistencies in choice of prescription drug insurance." *American Economic Review* 106 (8): 2145–2184.

Abrahamowicz, Michal, Yongling Xiao, Raluca Ionescu-Ittu, and Diane Lacaille. 2007. "Simulations showed that validation of database-derived diagnostic criteria based on a small subsample reduced bias." *Journal of Clinical Epidemiology* 60 (6): 600–609.

Amershi, Saleema, Max Chickering, Steven M. Drucker, Bongshin Lee, Patrice Simard, and Jina Suh. 2015. "ModelTracker: Redesigning Performance Analysis Tools for Machine Learning." In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15,* 337–346. New York, New York, USA: ACM Press.

Andrade, Susan E., Jerry H. Gurwitz, K. Arnold Chan, James G. Donahue, Arne Beck, Myde Boles, Diana S M Buist, et al. 2002. "Validation of diagnoses of peptic ulcers and bleeding from administrative databases: A multi-health maintenance organization study." *Journal of Clinical Epidemiology* 55 (3): 310–313.

Belloni, Alexandre, and Victor Chernozhukov. 2013. "Least squares after model selection in high-dimensional sparse models." *Bernoulli* 19 (2): 521–547.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2011. "Inference for High-Dimensional Sparse Econometric Models": 1–41. arXiv: 1201.0220.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "Inference on Treatment Effects after Selection among High-Dimensional Controls." *The Review of Economic Studies* 81 (2): 608–650.

Boser, Bernhard E., Isabelle M Guyon, and Vladimir N. Vapnik. 1992. “A training algorithm for optimal margin classifiers.” In *Proceedings of the fifth annual workshop on Computational learning theory,* 144–152. New York: ACM Press.

Bradley, Andrew P. 1997. “The use of the area under the ROC curve in the evaluation of machine learning algorithms.” *Pattern Recognition* 30 (7): 1145–1159.

Breiman, Leo. 1996. “Bagging Predictors.” *Machine Learning* 24:123–140.

Breiman, Leo. 2001. “Random Forests.” *Machine Learning* 45 (5): 5–32.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and regression trees.* London: Chapman & Hall/CRC.

Bullano, Michael F., Siddhesh Kamat, Vincent J. Willey, Suna Barlas, Douglas J. Watson, and Susan K. Brenneman. 2006. “Agreement Between Administrative Claims and the Medical Record in Identifying Patients With a Diagnosis of Hypertension.” *Medical Care* 44 (5): 486–490.

Candes, Emmanuel, and Terence Tao. 2007. “The Dantzig selector: Statistical estimation when p is much larger than n.” *The Annals of Statistics* 35 (6): 2313–2351.

Chan, An-Wen, Kinwah Fung, Jennifer M. Tran, Jessica Kitchen, Peter C. Austin, Martin A. Weinstock, and Paula A. Rochon. 2016. “Application of Recursive Partitioning to Derive and Validate a Claims-Based Algorithm for Identifying Keratinocyte Carcinoma (Non-melanoma Skin Cancer).” *JAMA Dermatology* 152 (10): 1122.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. “SMOTE: Synthetic Minority Over-sampling Technique.” *Journal of Artificial Intelligence Research* 16:321–357.

Chen, Tianqi, and Carlos Guestrin. 2016. “XGBoost: A Scalable Tree Boosting System”: 1–13. arXiv: 1603.02754.

Chen, Wei, Tie-Yan Liu, Yanyan Lan, Zhiming Ma, and Hang Li. 2009. "Ranking Measures and Loss Functions in Learning to Rank." In *NIPS '07: Proceedings of the 22nd International Conference on Neural Information Processing Systems,* 315–323. March.

Cheng, Ching-Lan, Yea-Huei Yang Kao, Swu-Jane Lin, Cheng-Han Lee, and Ming Liang Lai. 2011. "Validation of the National Health Insurance Research Database with ischemic stroke cases in Taiwan." *Pharmacoepidemiology and drug safety* 20 (3): 236–42.

Cheng, Ching-Lan, Cheng-Han Lee, Po-Sheng Chen, Yi-Heng Li, Swu-Jane Lin, and Yea-Huei Kao Yang. 2014. "Validation of Acute Myocardial Infarction Cases in the National Health Insurance Research Database in Taiwan." 24 (6): 500–507.

Collet, David. 1999. *Modelling Binary Data.* Second. Boca Raton, Florida: Chapman & Hall/CRC.

Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-vector networks." *Machine Learning* 20 (3): 273–297.

Cristianini, Nello, and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* New York: Cambridge University Press.

DeLong, Elizabeth R., David M. DeLong, and Daniel L. Clarke-Pearson. 1988. "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach." *Biometrics* 44 (3): 837–845.

Donoho, D. L., and M. Elad. 2003. "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization." *Proceedings of the National Academy of Sciences* 100 (5): 2197–2202.

Donoho, David L. 2006. "For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution." *Communications on Pure and Applied Mathematics* 59 (7): 907–934.

Einav, Liran, Amy Finkelstein, and Paul Schrimpf. 2015. "The Response of Drug Expenditure to Contract Design in Medicare Part D." *Quarterly Journal of Economics* 130 (2): 841–899.

Fawcett, Tom. 2006. "An introduction to ROC analysis." *Pattern Recognition Letters* 27 (8): 861–874.

Fisher, Ronald A. 1936. "The use of multiple measurements in taxonomic problems." *Annals of Eugenics* 7:179–188.

Fix, Evelyn, and J.L. Hodges. 1951. *Discriminatory Analysis - Nonparametric Discrimination: Consistency Properties.* Technical report. U.S. Air Force, School of Aviation Medicine, Randolph Field.

Freeman, Jean L., Dong Zhang, Daniel H. Freeman, and James S. Goodwin. 2000. "An approach to identifying incident breast cancer cases using Medicare claims data." *Journal of Clinical Epidemiology* 53 (6): 605–614.

Friedman, Jerome H. 1991. "Multivariate Adaptive Regression Splines." *The Annals of Statistics* 19 (1): 1–67.

Friedman, Jerome H. 2001. "Greedy function approximation: A gradient boosting machine." *The Annals of Statistics* 29 (5): 1189–1232.

Friedman, Jerome H. 2002. "Stochastic gradient boosting." *Computational Statistics & Data Analysis* 38 (4): 367–378.

Friedman, Jerome H., and Bogdan E. Popescu. 2003. *Importance Sampled Learning Ensembles.* Technical report. Department of Statistics, Stanford University.

Friedman, Jerome, Trevor Hastie, and Rob Tibshirani. 2010. "Regularization Paths for Generalized Linear Models via Coordinate Descent." *Journal of statistical software* 33 (1): 1–22.

Gold, Heather T., and Huong T. Do. 2007. "Evaluation of three algorithms to identify incident breast cancer in medicare claims data." *Health Services Research* 42 (5): 2056–2069.

Gorina, Yelena, and Ellen A. Kramarow. 2011. "Identifying chronic conditions in medicare claims data: Evaluating the chronic condition data warehouse algorithm." *Health Services Research* 46 (5): 1610–1627.

Greene, William H. 2012. *Econometric analysis.* 7th. Upper Saddle River: Prentice Hall.

Guyon, Isabelle, and André Elisseeff. 2003. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research* 3:1157–1182.

Hara, Konan, Jun Tomio, Thomas Svensson, Rika Ohkuma, Akiko Kishi Svensson, and Tsutomu Yamazaki. 2018. "Association measures of claims-based algorithms for common chronic conditions were assessed using regularly collected data in Japan." *Journal of Clinical Epidemiology* 99:84–95.

Hasan, Asad, Zhiyu Wang, and Alireza S. Mahani. 2016. "Fast Estimation of Multinomial Logit Models: R Package mnlogit." *Journal of Statistical Software* 75 (3).

Hastie, Trevor. 2018. *gam: Generalized Additive Models.* `https://cran.r-project.org/package=gam`.

Hastie, Trevor, Andreas Buja, and Robert Tibshirani. 1995. "Penalized Discriminant Analysis." *The Annals of Statistics* 23 (1): 73–102.

Hastie, Trevor, and Robert Tibshirani. 1986. "Generalized Additive Models." *Statistical Science* 1 (3): 297–310.

Hastie, Trevor, and Robert Tibshirani. 1990. *Generalized Additive Models.* London: Chapman & Hall/CRC.

Hastie, Trevor, Robert Tibshirani, and Andreas Buja. 1994. "Flexible Discriminant Analysis by Optimal Scoring." *Journal of the American Statistical Association* 89 (428): 1255–1270.

Hastie, Trevor, Robert Tibshirani, and Jerome H. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd. 745. New York: Springer.

Hastie, Trevor, Robert Tibshirani, Friedrich Leisch, Kurt Hornik, and Brian D. Ripley. 2017. *mda: Mixture and Flexible Discriminant Analysis.* `https://cran.r-project.org/package=mda`.

Hebert, Paul L., Linda S. Geiss, Edward F. Tierney, Michael M. Engelgau, Barbara P. Yawn, and A. Marshall McBean. 1999. "Identifying Persons with Diabetes Using Medicare Claims Data." *American Journal of Medical Quality* 14 (6): 270–277.

Helleputte, Thibault. 2017. *LiblineaR: Linear Predictive Models Based on the LIBLINEAR C/C++ Library*. `https://cran.r-project.org/package=LiblineaR`.

Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. "A Fast Learning Algorithm for Deep Belief Nets." *Neural Computation* 18 (7): 1527–1554.

Hoerl, Arthur E., and Robert W. Kennard. 1970. "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics* 12 (1): 55.

Iizuka, Toshiaki. 2012. "Physician agency and adoption of generic pharmaceuticals." *American Economic Review* 102 (6): 2826–2858.

Japan Atherosclerosis Society (Eds.) 2013. *Guidelines for the management of Dyslipidemia 2013 (in Japanese)*. Tokyo: Japan Atherosclerosis Society.

Katz, Jeffrey N., Jane Barrett, Matthew H. Liang, Anne M. Bacon, Herbert Kaplan, Raphael I. Kieval, Stephen M. Lindsey, et al. 1997. "Sensitivity and positive predictive value of medicare part B physician claims for rheumatologic diagnoses and procedures." *Arthritis & Rheumatism* 40 (9): 1594–1600.

Kawasumi, Yuko, Michal Abrahamowicz, Pierre Ernst, and Robyn Tamblyn. 2011. "Development and validation of a predictive algorithm to identify adult asthmatics from medical services and pharmacy claims databases." *Health Services Research* 46 (3): 939–963.

Kern, Elizabeth F. O., Miriam Maney, Donald R. Miller, Chin-Lin Tseng, Anjali Tiwari, Mangala Rajan, David Aron, and Leonard Pogach. 2006. "Failure of ICD-9-CM codes to identify patients with comorbid chronic kidney disease in diabetes." *Health Services Research* 41 (2): 564–580.

Khera, Rohan, Karen B. Dorsey, and Harlan M. Krumholz. 2018. "Transition to the ICD-10 in the United States." *JAMA* 320 (2): 133.

Kimura, Shinya, Toshihiko Sato, Shunya Ikeda, Mitsuhiko Noda, and Takeo Nakayama. 2010. "Development of a Database of Health Insurance Claims: Standardization of Disease Classifications and Anonymous Record Linkage." *Journal of Epidemiology* 20 (5): 413–419.

Klabunde, Carrie N., Linda C. Harlan, and Joan L. Warren. 2006. "Data sources for measuring comorbidity: a comparison of hospital records and medicare claims for cancer patients." *Medical care* 44 (10): 921–8.

Layton, J. Bradley, Yoonsang Kim, G. Caleb Alexander, and Sherry L. Emery. 2017. "Association Between Direct-to-Consumer Advertising and Testosterone Testing and Initiation in the United States, 2009-2013." *JAMA* 317 (11): 1159–1166.

LeCun, Y. 1989. *Generalization and Network Design Strategies.* Technical report. Department of Computer Science, Univ. of Toronto.

LeCun, Y., L. Bottou, Y. Bengio, and P. Haffner. 1998. "Gradient-based learning applied to document recognition." *Proceedings of the IEEE* 86 (11): 2278–2324.

Lin, Yi, and Hao Helen Zhang. 2006. "Component selection and smoothing in multivariate nonparametric regression." *Annals of Statistics* 34 (5): 2272–2297.

Losina, Elena, Jane Barrett, John A. Baron, and Jeffrey N. Katz. 2003. "Accuracy of Medicare claims data for rheumatologic diagnoses in total hip replacement recipients." *Journal of Clinical Epidemiology* 56 (6): 515–519.

Malley, J. D., J. Kruppa, A. Dasgupta, K. G. Malley, and A. Ziegler. 2012. "Probability Machines." *Methods of Information in Medicine* 51 (01): 74–81.

Mardia, Kanti V., John T. Kent, and John M. Bibby. 1979. *Multivariate Analysis.* New York: Academic Press.

McCulloch, Warren S., and Walter Pitts. 1943. "A logical calculus of the ideas immanent in nervous activity." *The Bulletin of Mathematical Biophysics* 5 (4): 115–133.

McWilliams, J. Michael, Laura A. Hatfield, Michael E. Chernew, Bruce E. Landon, and Aaron L. Schwartz. 2016. "Early Performance of Accountable Care Organizations in Medicare." *New England Journal of Medicine* 374 (24): 2357–2366.

Ministry of Health Labour and Welfare. 2012. *Survey on State of Employees' Health in 2012 (in Japanese).* Accessed October 10, 2018. `http://www.e-stat.go.jp/`.

Mitchell, Janet B., Thomas Bubolz, John E. Paul, Chris L. Pashos, José J. Escarce, Lawrence H. Muhlbaier, John M. Wiesman, Wanda W. Young, Robert S. Epstein, and Jonathan C. Javitt. 1994. "Using Medicare Claims for Outcomes Research." *Medical Care* 32 (7): JS38–JS51.

Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. 2012. *Foundations of Machine Learning*. Cambridge and London: MIT Press.

Morgan, James N., and John A. Sonquist. 1963. "Problems in the Analysis of Survey Data, and a Proposal." *Journal of the American Statistical Association* 58 (302): 415–434.

Muhajarine, Nazeem, Cameron Mustard, Leslie L. Roos, T. Kue Young, and Dale E. Gelskey. 1997. "Comparison of survey and physician claims data for detecting hypertension." *Journal of Clinical Epidemiology* 50 (6): 711–718.

Nattinger, Ann B., Purushottam W. Laud, Ruta Bajorunaite, Rodney A. Sparapani, and Jean L. Freeman. 2004. "An Algorithm for the Use of Medicare Claims Data to Identify Women with Incident Breast Cancer." *Health Services Research* 39 (6): 1733–1750.

Newton, K. M., P. L. Peissig, A. N. Kho, S. J. Bielinski, R. L. Berg, V. Choudhary, M. Basford, et al. 2013. "Validation of electronic medical record-based phenotyping algorithms: results and lessons learned from the eMERGE network." *Journal of the American Medical Informatics Association* 20:e147–e154.

Nordstrom, Beth L., Heather S. Norman, Timothy J. Dube, Marsha A. Wilcox, and Alexander M. Walker. 2007. "Identification of abacavir hypersensitivity reaction in health care claims data." *Pharmacoepidemiology and Drug Safety* 16 (3): 289–296.

Nuti, Sudhakar V., Li Qin, John S. Rumsfeld, Joseph S. Ross, Frederick A. Masoudi, Sharon-Lise T. Normand, Karthik Murugiah, Susannah M. Bernheim, Lisa G. Suter, and Harlan M. Krumholz. 2016. "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* 315 (6): 582–92.

Østbye, Truls, Donald H. Taylor, Elizabeth C. Clipp, Lynn Van Scoyoc, and Brenda L. Plassman. 2008. "Identification of dementia: Agreement among national survey data, medicare claims, and death certificates." *Health Services Research* 43 (1): 313–326.

Perkins, Neil J., and Enrique F. Schisterman. 2006. "The inconsistency of "optimal" cutpoints obtained using two criteria based on the receiver operating characteristic curve." *American Journal of Epidemiology* 163 (7): 670–675.

Pinto, David. 2018. *fastknn: Build Fast k-Nearest Neighbor Classifiers.* `https://github.com/davpinto/fastknn`.

Quam, Lois, Lynda B.M. Ellis, Pat Venus, Jon Clouse, Cynthia G. Taylor, and Sheila Leatherman. 1993. "Using claims data for epidemiologic research. The concordance of claims-based criteria with the medical record and patient survey for identifying a hypertensive population." *Medical care* 31 (6): 498–507.

Quan, Hude, Nadia Khan, Brenda R. Hemmelgarn, Karen Tu, Guanmin Chen, Norm Campbell, Michael D. Hill, William A. Ghali, and Finlay A. McAlister. 2009. "Validation of a case definition to define hypertension using administrative data." *Hypertension* 54 (6): 1423–1428.

R Core Team. 2018. *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing. `https://www.r-project.org/`.

Raskutti, Garvesh, Martin J. Wainwright, and Bin Yu. 2011. "Minimax Rates of Estimation for High-Dimensional Linear Regression Over Lq-Balls." *IEEE Transactions on Information Theory* 57 (10): 6976–6994.

Ravikumar, Pradeep, John Lafferty, Han Liu, and Larry Wasserman. 2009. "Sparse additive models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (5): 1009–1030.

Rector, Thomas S., Steven L. Wickstrom, Mona Shah, N. Thomas Greeenlee, Paula Rheault, Jeannette Rogowski, Vicki Freedman, John Adams, and José J. Escarce. 2004. "Specificity and sensitivity of claims-based algorithms for identifying members of Medicare+Choice health plans that have chronic medical conditions." *Health Services Research* 39 (6): 1839–1857.

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: an open-source package for R and S+ to analyze and compare ROC curves." *BMC Bioinformatics* 12 (1): 77.

Robinson, J. Reneé, T. Kue Young, Leslie L. Roos, and Dale E. Gelskey. 1997. "Estimating the burden of disease. Comparing administrative data and self-reports." *Medical care* 35 (9): 932–947.

Rong-En, Fan, Chang Kai-Wei, Hsieh Cho-Jui, Wang Xiang-Rui, and Lin Chih-Jen. 2008. "LIBLINEAR: A Library for Large Linear Classification." *Journal of Machine Learning Research* 9:1871–1874.

Rosenblatt, F. 1958. "The perceptron: A probabilistic model for information storage and organization in the brain." *Psychological Review* 65 (6): 386–408.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986a. "Learning Internal Representations by Error Propagation." Chap. 8 in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* edited by David E. Rumelhart and James L. McClelland, 1:318–362. Cambridge: MIT Press.

Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986b. "Learning representations by back-propagating errors." *Nature* 323 (6088): 533–536.

Sands, Kenneth, Gordon Vineyard, James Livingston, Cindy Christiansen, and Richard Platt. 1999. "Efficient Identification of Postdischarge Surgical Site Infections: Use of Automated Pharmacy Dispensing Information, Administrative Data, and Medical Record Information." *The Journal of Infectious Diseases* 179 (2): 434–441.

Schermerhorn, Marc L., Dominique B. Buck, A. James O'Malley, Thomas Curran, John C. McCallum, Jeremy Darling, and Bruce E. Landon. 2015. "Long-Term Outcomes of Abdominal Aortic Aneurysm in the Medicare Population." *New England Journal of Medicine* 373 (4): 328–338.

Scholes, D., O. Yu, M. A. Raebel, B. Trabert, and V. L. Holt. 2011. "Improving automated case finding for ectopic pregnancy using a classification algorithm." *Human Reproduction* 26 (11): 3163–3168.

Segal, Mark R. 2004. "Machine learning benchmarks and random forest regression." *UCSF: Center for Bioinformatics and Molecular Biostatistics.*

Shepard, Donald. 1968. "A two-dimensional interpolation function for irregularly-spaced data." In *Proceedings of the 23rd ACM National Conference,* 517–524. New York, USA: ACM.

Shevade, S. K., and S. S. Keerthi. 2003. "A simple and efficient algorithm for gene selection using sparse logistic regression." *Bioinformatics* 19 (17): 2246–2253.

Simard, Patrice, Yann LeCun, and John S. Denker. 1992. "Efficient pattern recognition using a new transformation distance." In *NIPS'92: Proceedings of the 5th International Conference on Neural Information Processing Systems,* 50–58.

Stevenson, Mark, Telmo Nunes, Cord Heuer, Jonathon Marshall, Javier Sanchez, Ron Thornton, Jeno Reiczigel, et al. 2018. *epiR: Tools for the Analysis of Epidemiological Data.* `https://cran.r-project.org/package=epiR`.

Taylor, Donald H., Gerda G. Fillenbaum, and Michael E. Ezell. 2002. "The accuracy of medicare claims data in identifying Alzheimer's disease." *Journal of Clinical Epidemiology* 55 (9): 929–937.

Taylor, Donald H., Truls Østbye, Kenneth M. Langa, David Weir, and Brenda L. Plassman. 2009. "The accuracy of medicare claims as an epidemiological tool: The case of dementia revisited." *Journal of Alzheimer's Disease* 17 (4): 807–815.

Tessier-Sherman, Baylah, Deron Galusha, Oyebode a Taiwo, Linda Cantley, Martin D Slade, Sharon R Kirsche, and Mark R Cullen. 2013. "Further validation that claims data are a useful tool for epidemiologic research on hypertension." *BMC public health* 13:51.

The Japan Diabetes Society (Eds.) 2016. *Guidelines for the management of Diabetes 2016 (in Japanese).* Tokyo: The Japan Diabetes Society.

The Japanese Society of Hypertension (Eds.) 2014. *Guidelines for the management of Hypertension 2014 (in Japanese).* Tokyo: The Japanese Society of Hypertension.

Tibshirani, Robert. 1996. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58 (1): 267–288.

Tikhonov, Andrey N. 1963. "Solution of incorrectly formulated problems and the regularization method." *Soviet Math. Dokl.* 4:1035–1038.

Tu, Karen, Doug Manuel, Kelvin Lam, Doug Kavanagh, Tezeta F. Mitiku, and Helen Guo. 2011. "Diabetics can be identified in an electronic medical record using laboratory tests and prescriptions." *Journal of Clinical Epidemiology* 64 (4): 431–435.

Upadhyaya, Sudhi G., Dennis H. Murphree, Che G. Ngufor, Alison M. Knight, Daniel J. Cronk, Robert R. Cima, Timothy B. Curry, Jyotishman Pathak, Rickey E. Carter, and Daryl J. Kor. 2017. "Automated Diabetes Case Identification Using Electronic Health Record Data at a Tertiary Care Facility." *Mayo Clinic Proceedings: Innovations, Quality & Outcomes* 1 (1): 100–110.

Van Walraven, Carl, Carol Bennett, and Alan J. Forster. 2011. "Administrative database research infrequently used validated diagnostic or procedural codes." *Journal of Clinical Epidemiology* 64 (10): 1054–1059.

Vapnik, Vladimir N. 1999. "An Overview of Statistical Learning Theory." *IEEE Transactions on Neural Networks* 10 (5): 988–999.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S.* Fourth. New York: Springer.

Virnig, Beth A, and Marshall McBean. 2001. "Administrative Data for Public Health Surveillance and Planning." *Annual Review of Public Health* 22 (1): 213–230.

Waldron, Levi, Melania Pintilie, Ming-sound Tsao, Frances A Shepherd, Curtis Huttenhower, and Igor Jurisica. 2011. "Optimized application of penalized regression methods to diverse genomic data." *Bioinformatics* 27 (24): 3399–3406.

Walraven, Carl van, and Ian Colman. 2016. "Migraineurs were reliably identified using administrative data." *Journal of Clinical Epidemiology* 71:68–75.

WHO. 2018a. *WHO - International Classification of Diseases.* Accessed October 10. `http://www.who.int/classifications/icd/en/`.

WHO. 2018b. *WHOCC - ATC/DDD Index.* Accessed October 10. `https://www.whocc.no/atc_ddd_index/`.

Wilchesky, Machelle, Robyn M. Tamblyn, and Allen Huang. 2004. "Validation of diagnostic codes within medical services claims." *Journal of Clinical Epidemiology* 57 (2): 131–141.

Wooldridge, Jeffrey M. 2010. *Econometric analysis of cross section and panel data.* 2nd. Cambridge: MIT Press.

Wright, Marvin N., and Andreas Ziegler. 2017. "ranger : A Fast Implementation of Random Forests for High Dimensional Data in C++ and R." *Journal of Statistical Software* 77 (1).

Yamana, Hayato, Hiromasa Horiguchi, Kiyohide Fushimi, and Hideo Yasunaga. 2016. "Comparison of Procedure-Based and Diagnosis-Based Identifications of Severe Sepsis and Disseminated Intravascular Coagulation in Administrative Data." *Journal of Epidemiology* 26 (10): 1–8.

Yamana, Hayato, Mutsuko Moriwaki, Hiromasa Horiguchi, Mariko Kodan, Kiyohide Fushimi, and Hide Yasunaga. 2017. "Validity of diagnoses, procedures, and laboratory data in Japanese administrative data." *Journal of Epidemiology:* 1–7.

Youden, W. J. 1950. “Index for rating diagnostic tests.” *Cancer* 3 (1): 32–35.

Zhu, JI, and Trevor Hastie. 2004. “Classification of gene microarrays by penalized logistic regression.” *Biostatistics* 5 (3): 427–443.

Zou, Hui, and Trevor Hastie. 2005. “Regularization and variable selection via the elastic net.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2): 301–320.

List of Figures

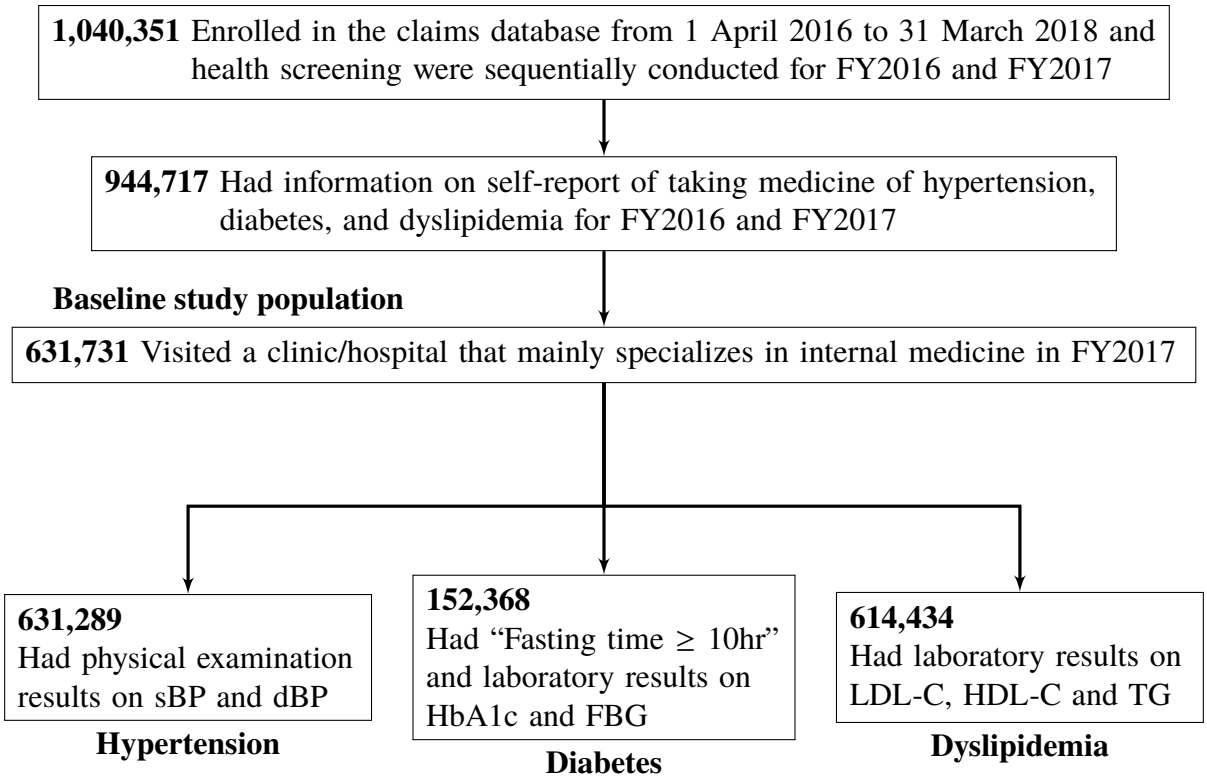

Figure 1. Flowchart of inclusion and exclusion of study participants

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.

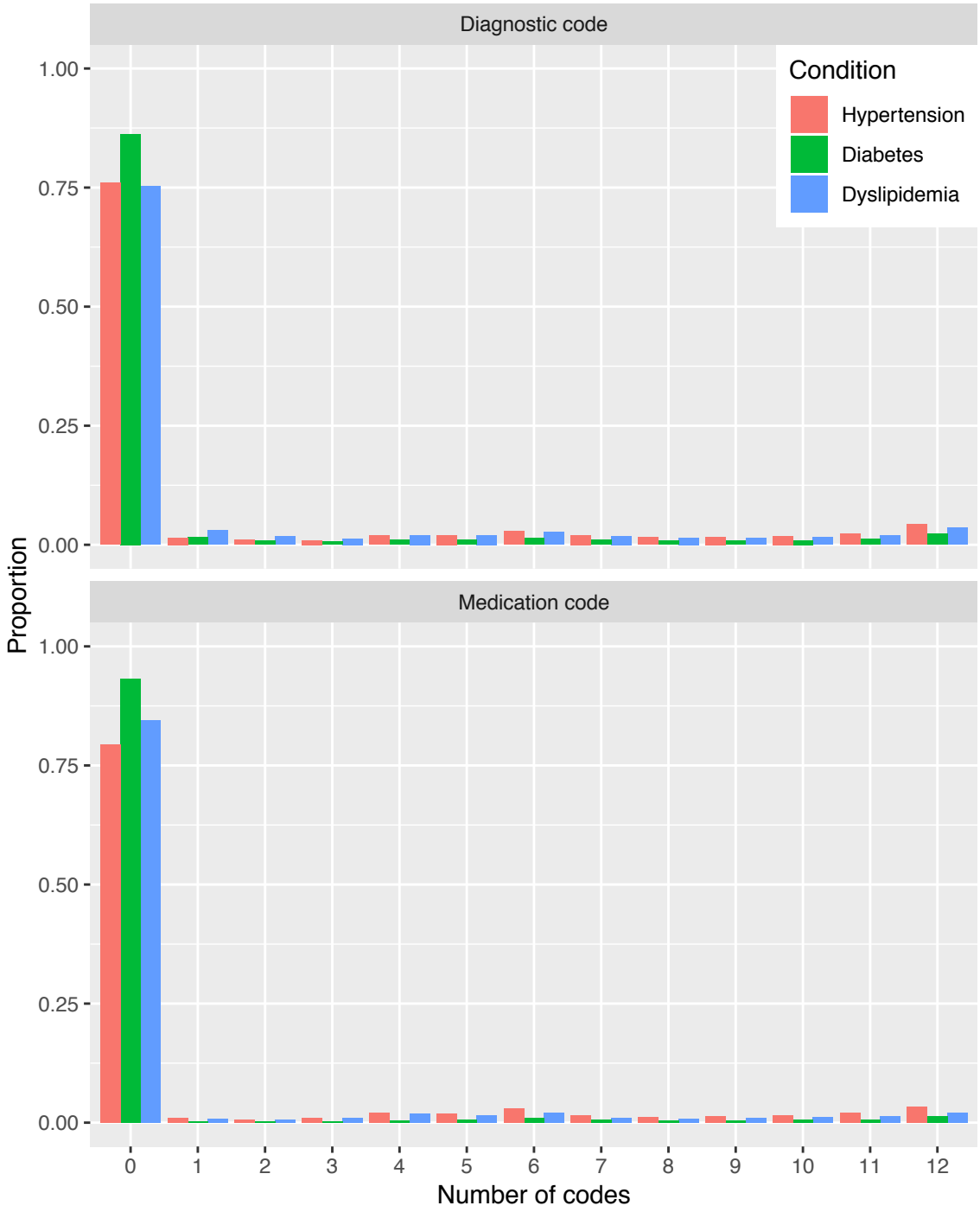

Figure 2. Probability mass function of the number of observations of diagnostic/medication code corresponding to hypertension, diabetes, and dyslipidemia for the baseline study population

The horizontal axis is the number of observations of diagnostic/medication code corresponding to hypertension, diabetes, or dyslipidemia. Observations of diagnostic or medication codes on claims were counted as one occurrence when information was accrued from the same month. The vertical axis shows the proportion of enrollees whose claims contain the designated number of observations of diagnostic/medication code corresponding to hypertension, diabetes, or dyslipidemia in FY2017.

The diagnostic codes corresponding to hypertension, diabetes, and dyslipidemia were, respectively, defined as the International Classification of Diseases and Related Health Problems, tenth revision (ICD-10) code I10-I15, E10-E14, and E78. The medication codes corresponding to hypertension, diabetes, and dyslipidemia were, respectively, defined as the World Health Organization-anatomical therapeutic chemical (WHO-ATC) code C08 (calcium channel blockers) and C09 (agents acting on the renin-angiotensin system), A10 (drugs used in diabetes), and C10 (lipid-modifying agents).

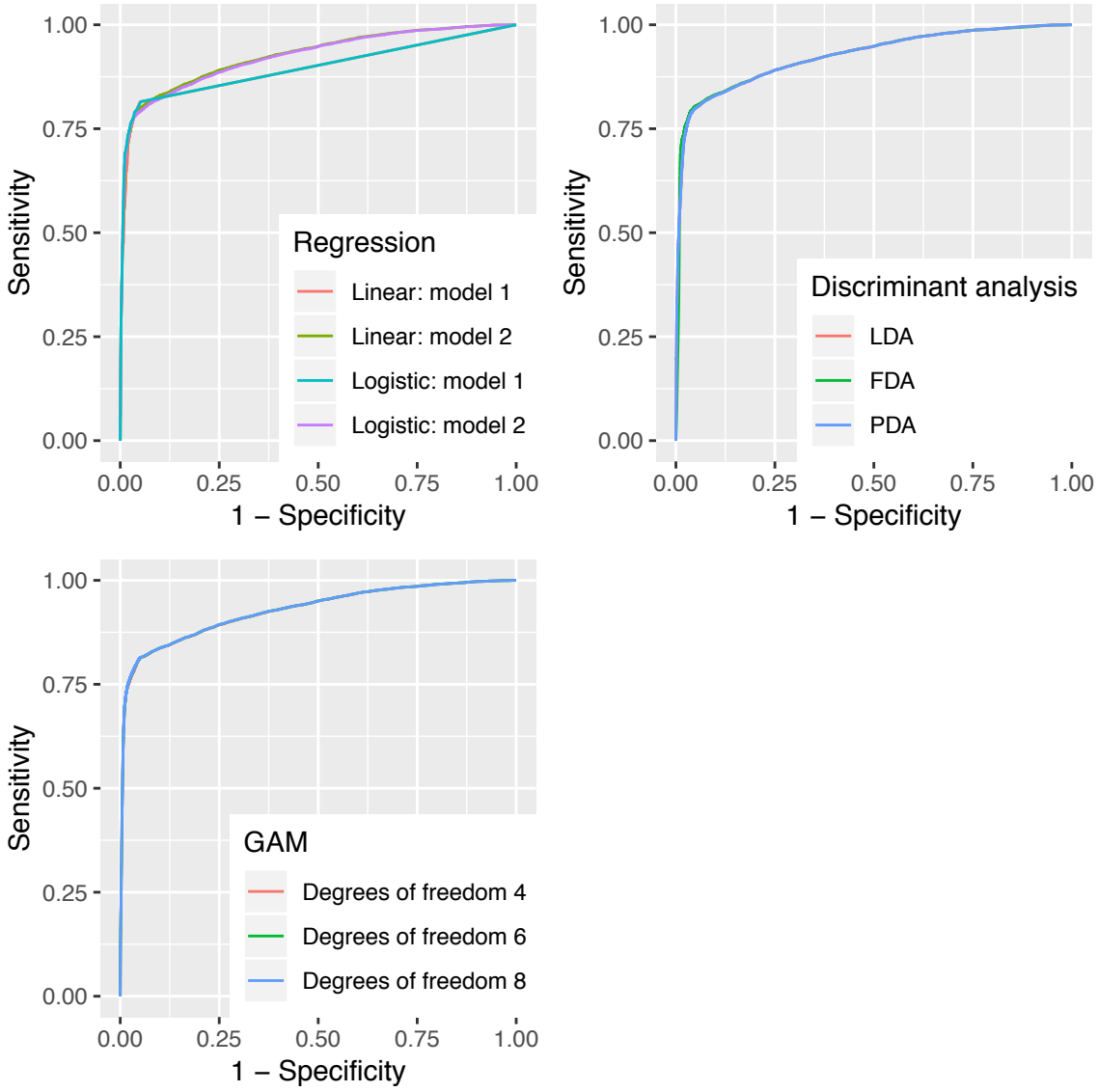

Figure 3A. Receiver operating characteristic curve for claims-based algorithms derived from statistical methods with a condition-specific variable selection according to hypertension

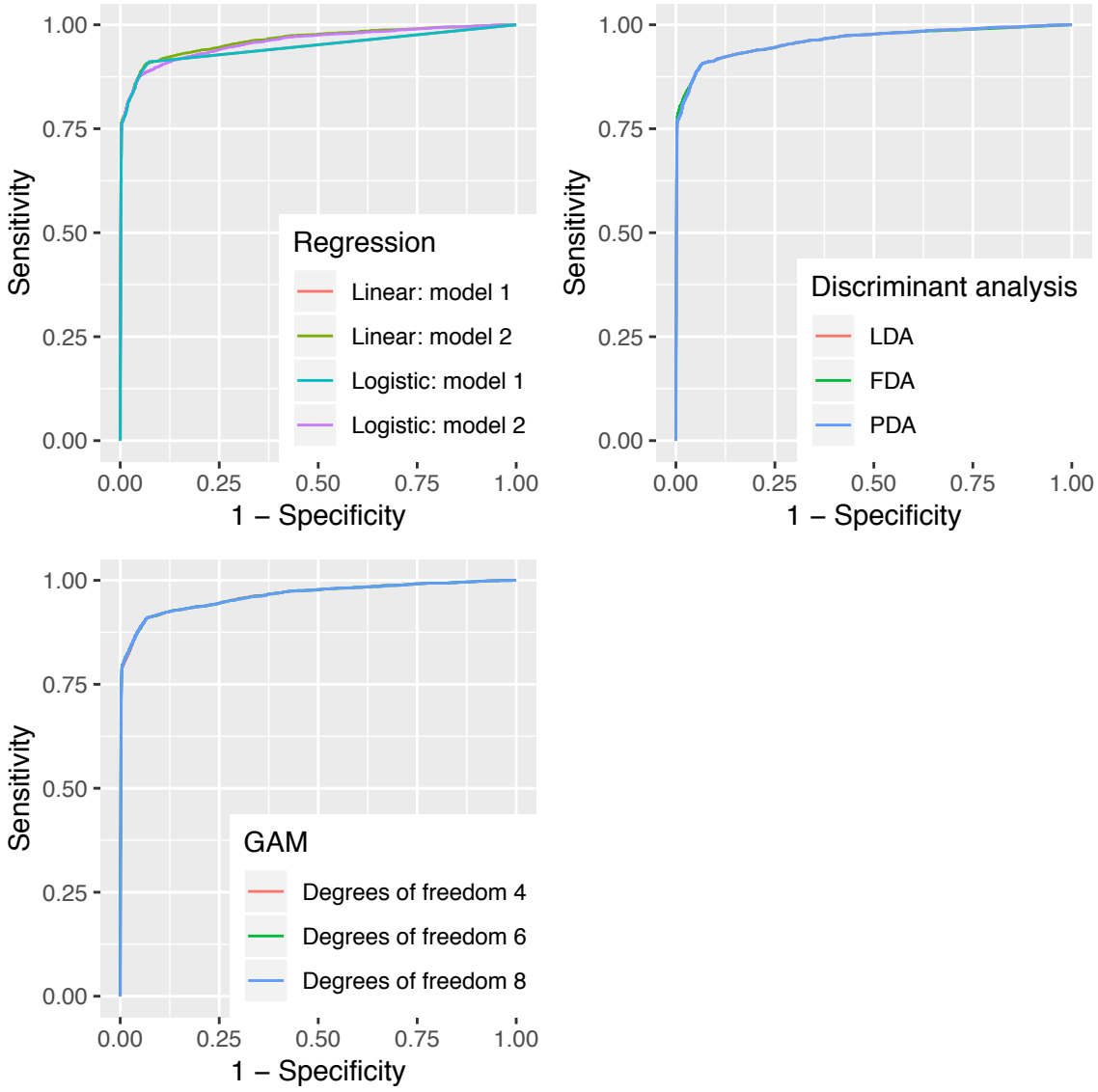

Figure 3B. Receiver operating characteristic curve for claims-based algorithms derived from statistical methods with a condition-specific variable selection according to diabetes

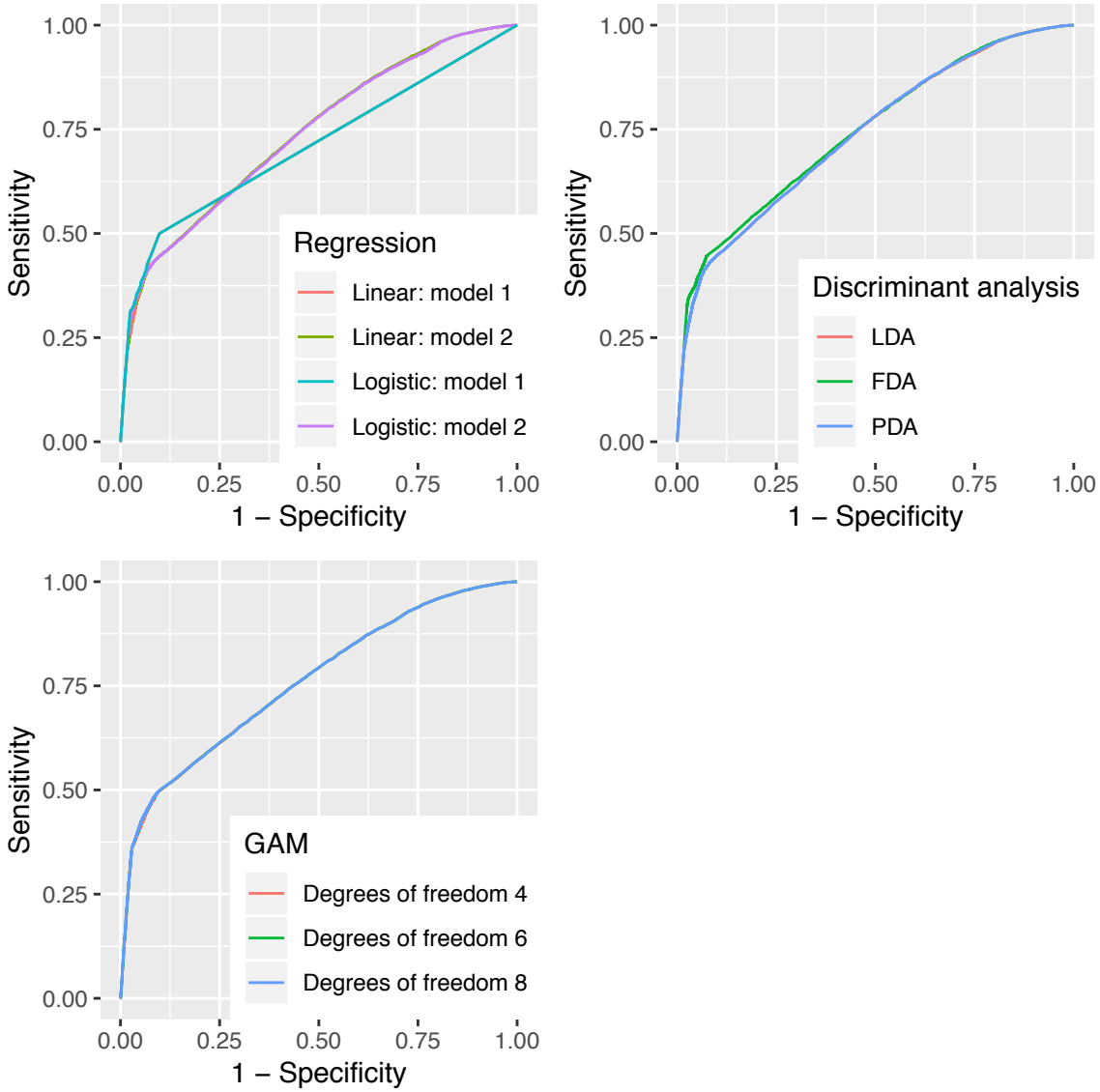

Figure 3C. Receiver operating characteristic curve for claims-based algorithms derived from statistical methods with a condition-specific variable selection according to dyslipidemia

Abbreviations: FDA, flexible discriminant analysis; GAM, generalized additive model; LDA, linear discriminant analysis; PDA, penalized discriminant analysis.

The number of observations of each of the diagnostic code and the medication code was used as input variables for model 1 in regression. Age and gender were added to these input variables for the other models.

Three types of discriminant analysis were conducted: linear discriminant analysis (LDA); flexible discriminant analysis (FDA); and penalized discriminant analysis (PDA). Multivariate adaptive regression splines with the second degree of interaction were used as the basis function in the FDA. In the PDA, a linear basis with the L_2-penalty was used, and the regularization coefficient was determined by cross-validation.

The additive logistic regression model with a cubic smoothing spline for each input variable except gender was used in the generalized additive model (GAM). The hyperparameter, the degrees of freedom for each smoothing spline, was set to the same value for all splines. I selected three values for the degrees of freedom, four, six, and eight, to generate three GAM prediction functions.

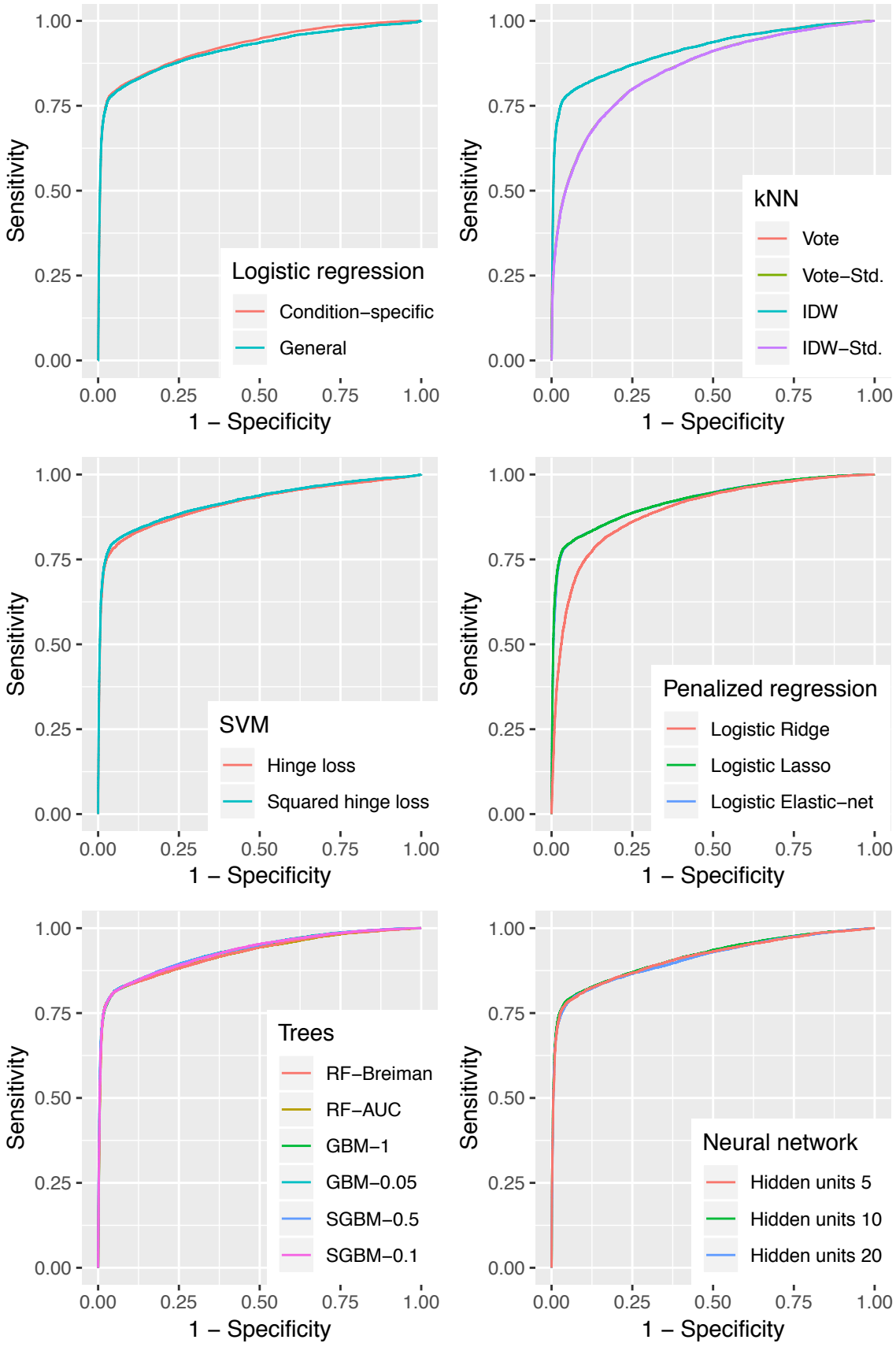

Figure 4A. Receiver operating characteristic curve for claims-based algorithms derived from statistical methods without a condition-specific variable selection according to hypertension

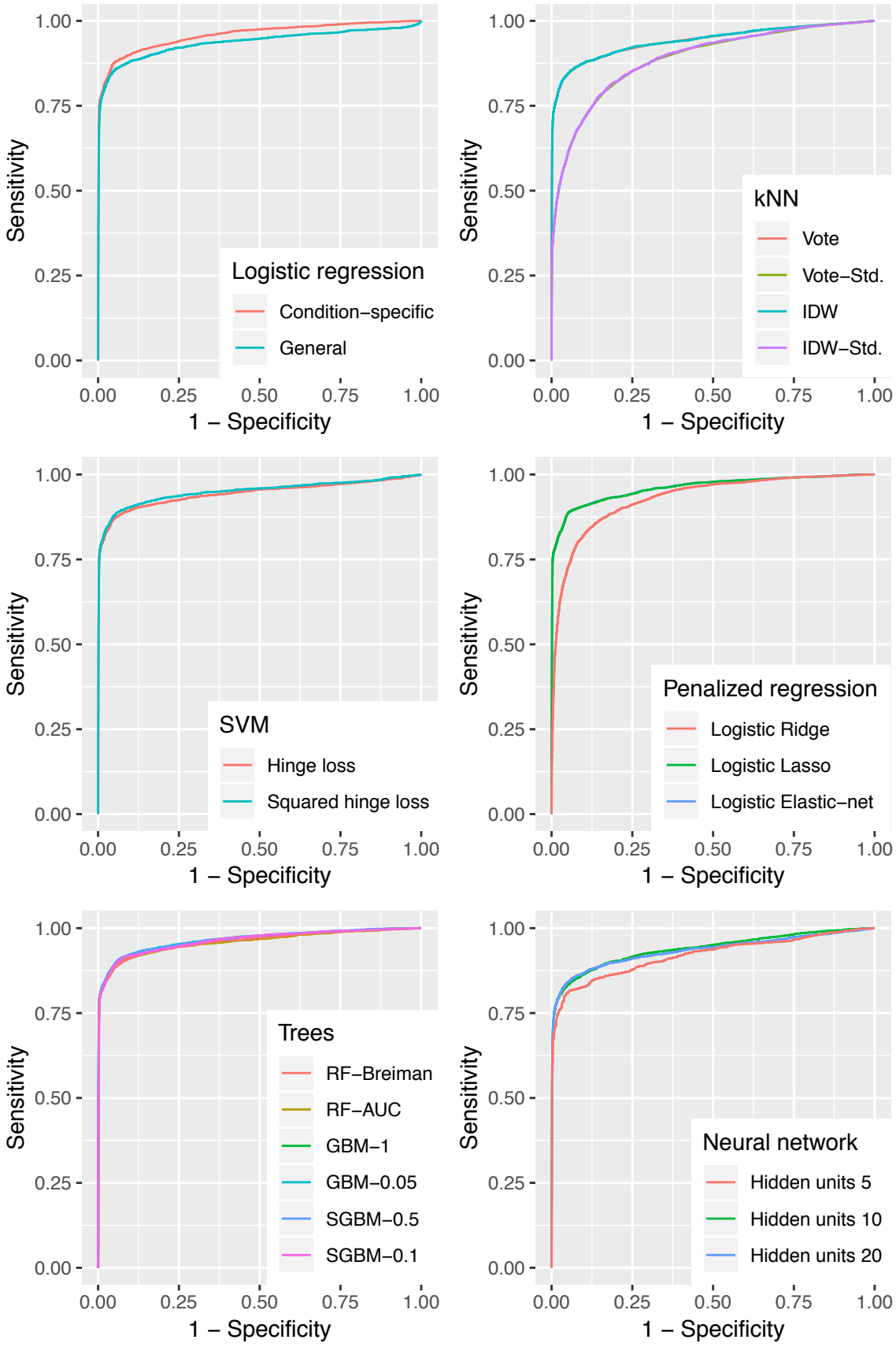

Figure 4B. Receiver operating characteristic curve for claims-based algorithms derived from statistical methods without a condition-specific variable selection according to diabetes

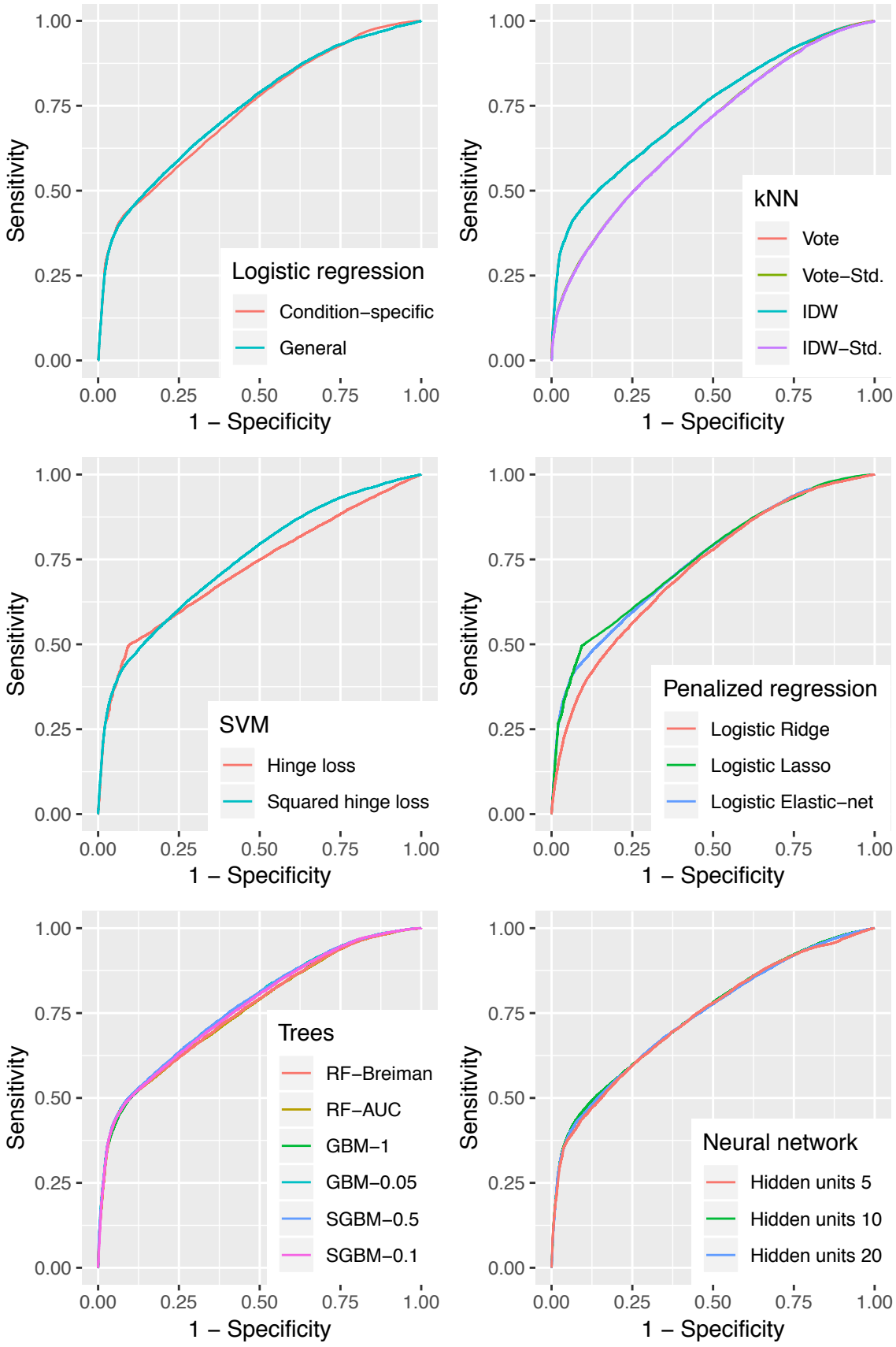

Figure 4C. Receiver operating characteristic curve for claims-based algorithms derived from statistical methods without a condition-specific variable selection according to dyslipidemia

Abbreviations: AUC, area under the receiver operating characteristic curve; GBM, gradient boosting machine; IDW, inverse distance weighting; ISLE, importance sampled learning ensemble; kNN, k-nearest neighbor; SGBM, stochastic gradient boosting machine; Std., standardized; SVM, support vector machine; RF, random forest.

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 condition-specific in the logistic regression. Condition-specific in the logistic regression is the same as model 2 of the logistic regression. General in the logistic regression fitted a logistic regression model to the dataset that was trimmed properly.

The Euclidean distance with raw or standardized (i.e., re-scaled 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), the L_1-penalty (logistic lasso), and the elastic-net penalty (logistic elastic-net) were applied. The regularization coefficient and the elastic-net mixing parameter were determined by cross-validation.

Two types of tree-based models were applied: the random forest and the importance sampled learning ensemble (ISLE). The minimum node size was set to ten for each tree, and two-hundred trees were bagged in the random forest. The number of variables selected for each split was first set to the value recommended by Breiman, $\lfloor\sqrt{p}\rfloor$, where p is the number of input variables, and then tuned using the validation set next. Denote the ISLE with the subsampling ratio equals to one and less than one by ISLE-gradient boosting machine (GBM) and ISLE-stochastic gradient boosting machine (SGBM), respectively. I fixed the depth to be six for all ISLEs. For the ISLE-GBM, I selected the learning rate to be 1 and 0.05. For the ISLE-SGBM, I fixed the learning rate to be 0.1 and selected the subsampling ratio to be 0.5 and 0.1. The number of trees and the 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: five, ten, and twenty. All hidden units were fully connected with the nodes in the input and output layer. Weight decay was employed for the regularization of the parameters, and the regularization coefficient of it was tuned using the validation set.

List of Tables

Table 1. Summary statistics of enrollees' characteristics and health screening results for each fiscal year

	FY2016		FY2017	
Variables	Mean	SD	Mean	SD
Demographics				
Male	–	–	0.80	–
Age* (year)	–	–	48.0	10.4
Visited clinic/hospital				
Any clinic/hospital†	0.85	–	0.85	–
Primary care clinic/hospital‡	0.67	–	0.67	–
Health screening results				
Fasting time ≥ 10 hours§	0.81	–	0.81	–
Systolic blood pressure (mmHg)	121.5	15.8	122.1	15.9
Diastolic blood pressure (mmHg)	75.5	11.7	75.9	11.8
Fasting blood glucose (mg/dL)	96.7	18.5	97.3	19.0
Hemoglobin A1c (%)	5.56	0.64	5.59	0.64
Low-density lipoprotein cholesterol (mg/dL)	121.1	30.8	121.3	30.6
High-density lipoprotein cholesterol (mg/dL)	60.6	15.9	60.9	16.1
Triglyceride (mg/dL)	117.1	94.0	118.3	94.5
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.

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.

* 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.

Table 2. Cumulative distribution of the proportion of enrollees whose claims contain the ICD-10/WHO-ATC code at least once in the baseline study population

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

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

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 baseline study population. Cumulative 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.

Table 3A. Association measures and their 95% confidence intervals for claims-based algorithms derived from conventional methods according to hypertension (N = 157,822, Prevalence = 25.4%)

Claims-based algorithm	Sensitivity			Specificity			PPV			NPV		
	%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Diagnostic code												
Baseline	80.4	80.0	80.8	95.1	95.0	95.2	84.9	84.5	85.3	93.4	93.3	93.6
Threshold = 2	77.7	77.3	78.1	96.2	96.1	96.3	87.4	87.0	87.7	92.7	92.5	92.8
Threshold = 3	75.5	75.1	76.0	96.8	96.7	96.9	88.9	88.6	89.2	92.1	91.9	92.2
Medication code												
Baseline	74.8	74.4	75.3	97.8	97.7	97.8	91.9	91.6	92.2	91.9	91.8	92.1
Threshold = 2	73.0	72.5	73.4	98.3	98.2	98.4	93.6	93.4	93.9	91.4	91.3	91.6
Threshold = 3	71.0	70.6	71.5	98.6	98.5	98.7	94.5	94.3	94.8	90.9	90.7	91.0
Alternative medication code	76.7	76.3	77.2	96.7	96.6	96.8	88.7	88.4	89.1	92.4	92.3	92.6
Combined												
Baseline	74.4	73.9	74.8	98.1	98.0	98.2	93.1	92.8	93.4	91.8	91.7	92.0
Threshold = 2	72.5	72.1	73.0	98.5	98.4	98.6	94.3	94.0	94.6	91.3	91.2	91.5
Threshold = 3	70.5	70.1	71.0	98.8	98.7	98.8	95.1	94.8	95.3	90.8	90.6	90.9
Alternative medication code	75.9	75.5	76.3	97.7	97.6	97.8	91.8	91.5	92.1	92.2	92.1	92.4

Table 3B. Association measures and their 95% confidence intervals for claims-based algorithms derived from conventional methods according to diabetes (N = 38,092, Prevalence = 8.3%)

	Sensitivity			Specificity			PPV			NPV		
Claims-based algorithm	%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Diagnostic code												
Baseline	91.1	90.0	92.0	92.8	92.6	93.1	53.4	52.0	54.7	99.1	99.0	99.2
Threshold = 2	88.4	87.2	89.5	94.3	94.0	94.5	58.1	56.7	59.5	98.9	98.8	99.0
Threshold = 3	86.3	85.0	87.5	95.1	94.9	95.3	61.3	59.8	62.7	98.7	98.6	98.8
Medication code												
Baseline	79.2	77.7	80.6	99.5	99.5	99.6	93.9	92.9	94.8	98.2	98.0	98.3
Threshold = 2	77.2	75.7	78.7	99.7	99.6	99.7	95.2	94.3	96.0	98.0	97.8	98.1
Threshold = 3	75.5	74.0	77.0	99.7	99.6	99.7	95.6	94.7	96.4	97.8	97.7	98.0
Combined												
Baseline	79.2	77.7	80.6	99.6	99.5	99.7	94.7	93.8	95.5	98.2	98.0	98.3
Threshold = 2	77.2	75.6	78.6	99.7	99.6	99.7	95.5	94.6	96.3	98.0	97.8	98.1
Threshold = 3	75.5	73.9	77.0	99.7	99.6	99.7	95.7	94.8	96.5	97.8	97.7	98.0

Table 3C. Association measures and their 95% confidence intervals for claims-based algorithms derived from conventional methods according to dyslipidemia (N = 153,608, Prevalence = 38.7%)

	Sensitivity			Specificity			PPV			NPV		
Claims-based algorithm	%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Diagnostic code												
Baseline	49.2	48.8	49.6	90.1	89.9	90.2	75.8	75.3	76.2	73.7	73.5	74.0
Threshold = 2	44.1	43.7	44.5	91.8	91.7	92.0	77.3	76.9	77.8	72.2	72.0	72.5
Threshold = 3	41.0	40.6	41.4	92.8	92.6	93.0	78.2	77.8	78.7	71.3	71.1	71.6
Medication code												
Baseline	35.8	35.5	36.2	96.9	96.8	97.0	88.0	87.6	88.4	70.5	70.2	70.7
Threshold = 2	34.0	33.7	34.4	97.1	97.0	97.2	88.3	87.8	88.7	70.0	69.7	70.2
Threshold = 3	32.5	32.1	32.9	97.3	97.2	97.4	88.5	88.1	88.9	69.5	69.3	69.8
Combined												
Baseline	35.8	35.4	36.1	97.0	96.9	97.1	88.2	87.8	88.6	70.5	70.2	70.7
Threshold = 2	34.0	33.6	34.3	97.2	97.1	97.3	88.4	87.9	88.8	69.9	69.7	70.2
Threshold = 3	32.4	32.0	32.8	97.4	97.3	97.5	88.6	88.2	89.0	69.5	69.2	69.7

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

At baseline, 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 CBA); (2) the medication code corresponding to the condition is found in the claims at least once (medication code-based CBA), and (3) the diagnostic code and the medication code corresponding to the condition are both found in the claims data at least once (combined CBA). Next, I changed the threshold for each algorithm by requiring the presence of two or three diagnostic and/or medication codes to consider a patient positive. In alternative medication code for hypertension, I broadened the definition of medication codes to include C02 (antihypertensive drugs), C03 (diuretic drugs), and C07 (beta-blocking agents). As the test set (25% of the baseline study population of each condition) was used in the calculation of the association measures, the sample size was 157,822, 38,092, and 153,608 for hypertension, diabetes, and dyslipidemia, respectively. For them, the prevalence which was determined by the gold standard for each condition was 25.4%, 8.3%, and 38.7% for hypertension, diabetes, and dyslipidemia, respectively. I calculated 95% CIs for all estimates of sensitivity, specificity, PPV, and NPV using exact binomial confidence limits.

Table 4A. Sensitivity analysis for the association measures of baseline claims-based algorithm according to hypertension

Claims-based algorithm	N	Prev. %	Sensitivity			Specificity			PPV			NPV		
			%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Diagnostic code														
Baseline	157,822	25.4	80.4	80.0	80.8	95.1	95.0	95.2	84.9	84.5	85.3	93.4	93.3	93.6
Population: clinic/hospital	200,678	22.4	76.4	76.1	76.8	96.1	96.0	96.2	84.9	84.6	85.3	93.4	93.3	93.5
Population: all	236,009	20.6	70.9	70.5	71.4	96.7	96.7	96.8	84.9	84.6	85.3	92.8	92.7	92.9
Population: age ≥ 50 years	77,139	39.4	84.2	83.8	84.6	91.9	91.7	92.1	87.1	86.7	87.5	89.9	89.7	90.2
Alternative gold standard	157,822	34.0	64.0	63.6	64.4	96.5	96.4	96.6	90.5	90.2	90.8	83.9	83.7	84.1
Medication code														
Baseline	157,822	25.4	74.8	74.4	75.3	97.8	97.7	97.8	91.9	91.6	92.2	91.9	91.8	92.1
Population: clinic/hospital	200,678	22.4	71.1	70.7	71.5	98.2	98.1	98.3	92.0	91.7	92.3	92.2	92.0	92.3
Population: all	236,009	20.6	66.0	65.6	66.4	98.5	98.5	98.6	92.0	91.7	92.3	91.8	91.7	91.9
Population: age ≥ 50 years	77,139	39.4	79.1	78.7	79.6	96.1	95.9	96.3	92.9	92.6	93.3	87.6	87.3	87.9
Alternative gold standard	157,822	34.0	58.0	57.6	58.5	98.5	98.5	98.6	95.3	95.1	95.5	82.0	81.8	82.2
Combined														
Baseline	157,822	25.4	74.4	73.9	74.8	98.1	98.0	98.2	93.1	92.8	93.4	91.8	91.7	92.0
Population: clinic/hospital	200,678	22.4	70.7	70.2	71.1	98.5	98.4	98.6	93.2	92.9	93.5	92.1	91.9	92.2
Population: all	236,009	20.6	65.6	65.2	66.0	98.8	98.7	98.8	93.2	92.9	93.5	91.7	91.6	91.9
Population: age ≥ 50 years	77,139	39.4	78.6	78.1	79.1	96.7	96.5	96.9	93.9	93.6	94.2	87.4	87.1	87.7
Alternative gold standard	157,822	34.0	57.5	57.1	57.9	98.9	98.8	98.9	96.3	96.1	96.5	81.9	81.6	82.1

Table 4B. Sensitivity analysis for the association measures of baseline claims-based algorithm according to diabetes

Claims-based algorithm	N	Prev. %	Sensitivity			Specificity			PPV			NPV		
			%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Diagnostic code														
Baseline	38,092	8.3	91.1	90.0	92.0	92.8	92.6	93.1	53.4	52.0	54.7	99.1	99.0	99.2
Population: clinic/hospital	47,903	7.1	89.1	88.0	90.1	94.0	93.8	94.3	53.2	51.9	54.5	99.1	99.0	99.2
Population: all	55,637	6.4	84.9	83.6	86.0	94.9	94.7	95.1	53.2	51.9	54.5	98.9	98.8	99.0
Population: age ≥ 50 years	19,484	12.9	91.6	90.4	92.6	89.1	88.6	89.5	55.3	53.8	56.8	98.6	98.4	98.8
Alternative gold standard	38,092	9.1	87.4	86.3	88.5	93.3	93.0	93.5	56.6	55.2	57.9	98.7	98.5	98.8
Medication code														
Baseline	38,092	8.3	79.2	77.7	80.6	99.5	99.5	99.6	93.9	92.9	94.8	98.2	98.0	98.3
Population: clinic/hospital	47,903	7.1	77.4	76.0	78.8	99.6	99.6	99.7	93.8	92.9	94.7	98.3	98.2	98.4
Population: all	55,637	6.4	73.8	72.3	75.2	99.7	99.6	99.7	93.8	92.9	94.7	98.2	98.1	98.3
Population: age ≥ 50 years	19,484	12.9	79.5	77.9	81.1	99.3	99.2	99.4	94.3	93.3	95.3	97.0	96.8	97.3
Alternative gold standard	38,092	9.1	73.1	71.6	74.6	99.7	99.6	99.7	95.7	94.8	96.4	97.4	97.2	97.5
Combined														
Baseline	38,092	8.3	79.2	77.7	80.6	99.6	99.5	99.7	94.7	93.8	95.5	98.2	98.0	98.3
Population: clinic/hospital	47,903	7.1	77.4	76.0	78.8	99.7	99.6	99.7	94.6	93.7	95.4	98.3	98.2	98.4
Population: all	55,637	6.4	73.7	72.2	75.2	99.7	99.7	99.8	94.6	93.7	95.4	98.2	98.1	98.3
Population: age ≥ 50 years	19,484	12.9	79.5	77.9	81.1	99.4	99.3	99.5	95.0	94.0	95.9	97.0	96.8	97.3
Alternative gold standard	38,092	9.1	73.0	71.5	74.5	99.7	99.7	99.8	96.4	95.6	97.1	97.4	97.2	97.5

Table 4C. Sensitivity analysis for the association measures of baseline claims-based algorithm according to dyslipidemia

Claims-based algorithm	N	Prev. %	Sensitivity			Specificity			PPV			NPV		
			%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Diagnostic code														
Baseline	153,608	38.7	49.2	48.8	49.6	90.1	89.9	90.2	75.8	75.3	76.2	73.7	73.5	74.0
Population: clinic/hospital	194,598	37.0	43.6	43.2	44.0	91.8	91.7	92.0	75.8	75.4	76.2	73.5	73.3	73.7
Population: all	228,048	36.0	38.2	37.9	38.5	93.1	93.0	93.3	75.8	75.4	76.2	72.8	72.6	73.0
Population: age ≥ 50 years	77,148	47.0	57.7	57.2	58.2	84.4	84.1	84.8	76.7	76.2	77.2	69.2	68.8	69.6
Alternative gold standard	153,608	57.7	38.4	38.1	38.7	92.9	92.7	93.1	88.0	87.7	88.4	52.5	52.3	52.8
Medication code														
Baseline	153,608	38.7	35.8	35.5	36.2	96.9	96.8	97.0	88.0	87.6	88.4	70.5	70.2	70.7
Population: clinic/hospital	194,598	37.0	31.7	31.3	32.0	97.5	97.4	97.5	87.9	87.5	88.3	70.8	70.6	71.1
Population: all	228,048	36.0	27.7	27.4	28.1	97.9	97.8	97.9	87.9	87.5	88.3	70.6	70.4	70.8
Population: age ≥ 50 years	77,148	47.0	44.7	44.2	45.2	94.5	94.3	94.8	87.9	87.4	88.4	65.8	65.4	66.2
Alternative gold standard	153,608	57.7	25.7	25.5	26.0	97.8	97.7	97.9	94.1	93.8	94.4	49.2	48.9	49.4
Combined														
Baseline	153,608	38.7	35.8	35.4	36.1	97.0	96.9	97.1	88.2	87.8	88.6	70.5	70.2	70.7
Population: clinic/hospital	194,598	37.0	31.6	31.2	31.9	97.5	97.4	97.6	88.1	87.7	88.5	70.8	70.6	71.1
Population: all	228,048	36.0	27.7	27.4	28.0	97.9	97.8	98.0	88.1	87.7	88.5	70.6	70.4	70.8
Population: age ≥ 50 years	77,148	47.0	44.6	44.1	45.1	94.6	94.4	94.8	88.0	87.5	88.5	65.8	65.4	66.2
Alternative gold standard	153,608	57.7	25.7	25.4	26.0	97.9	97.7	98.0	94.2	93.9	94.5	49.1	48.9	49.4

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

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 CBA); (2) the medication code corresponding to the condition is found in the claims at least once (medication code-based CBA), and (3) the diagnostic code and the medication code corresponding to the condition are both found in the claims data at least once (combined CBA). In addition to the baseline study population that was restricted to those who at least once in the fiscal year had visited a clinic/hospital that mainly specializes in internal medicine, the following study populations were considered: (1) enrollees who had visited any clinic/hospital at least once in FY2017 (population: clinic/hospital); (2) all enrollees including those who had not visited any clinic/hospital in FY2017 (population: all); and (3) enrollees aged 50 years or older in the baseline study population (population: age ≥ 50 years). I relaxed the criteria of the gold standard for each condition in the alternative gold standard. The prevalence was determined by the gold standard for the test set of each sensitivity analysis. I calculated 95% CIs for all estimates of sensitivity, specificity, PPV, and NPV using exact binomial confidence limits.

Table 5A. Association measures and their 95% confidence intervals for claims-based algorithms derived from statistical methods with a condition-specific variable selection according to hypertension

Method	AUC			Sensitivity			Specificity			PPV			NPV		
		95% CI		%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Regression															
Linear: model 1	0.895	0.891	0.899	81.4	80.7	82.2	95.0	94.7	95.2	84.8	84.1	85.5	93.7	93.4	93.9
Linear: model 2	0.925	0.922	0.929	79.5	78.6	80.4	95.5	95.0	96.0	85.9	84.7	87.2	93.1	92.8	93.4
Logistic: model 1	0.897	0.893	0.901	81.5	80.7	82.3	94.8	94.6	95.1	84.5	83.9	85.1	93.7	93.5	93.9
Logistic: model 2	0.924	0.920	0.927	78.3	77.3	79.2	96.2	95.6	96.7	87.6	86.0	89.0	92.8	92.5	93.0
Discriminant analysis															
Linear	0.925	0.922	0.929	79.6	78.7	80.4	95.5	95.0	96.1	86.0	84.5	87.4	93.1	92.9	93.4
Flexible	0.925	0.921	0.928	80.2	78.7	81.0	95.5	95.1	96.4	85.9	84.9	88.4	93.3	92.9	93.6
Penalized	0.925	0.922	0.929	79.6	78.7	80.5	95.5	94.9	96.0	85.9	84.5	87.2	93.1	92.9	93.4
Generalized additive model															
Degrees of freedom 4	0.928	0.925	0.932	81.3	80.5	82.1	95.1	94.8	95.5	85.2	84.3	86.2	93.6	93.4	93.9
Degrees of freedom 6	0.928	0.925	0.932	81.3	80.6	81.9	95.1	94.8	95.5	85.0	84.3	86.1	93.7	93.4	93.9
Degrees of freedom 8	0.929	0.925	0.932	81.4	80.4	82.3	95.1	94.8	95.4	85.0	84.3	86.0	93.7	93.4	94.0

Table 5B. Association measures and their 95% confidence intervals for claims-based algorithms derived from statistical methods with a condition-specific variable selection according to diabetes

Method	AUC			Sensitivity			Specificity			PPV			NPV		
		95% CI		%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Regression															
Linear: model 1	0.947	0.941	0.952	90.9	89.4	92.1	92.8	92.6	94.2	53.3	52.4	58.4	99.1	99.0	99.2
Linear: model 2	0.962	0.958	0.966	90.6	89.4	91.5	93.6	93.0	94.0	56.1	54.1	57.8	99.1	99.0	99.2
Logistic: model 1	0.946	0.941	0.952	91.0	89.6	92.1	92.8	92.5	94.3	53.3	52.4	58.8	99.1	99.0	99.2
Logistic: model 2	0.958	0.953	0.963	87.5	86.3	88.8	95.3	94.7	95.8	62.6	59.7	65.0	98.8	98.7	98.9
Discriminant analysis															
Linear	0.962	0.958	0.966	90.5	89.4	91.7	93.7	93.1	94.1	56.2	54.3	58.1	99.1	99.0	99.2
Flexible	0.962	0.957	0.966	90.4	89.4	91.5	93.5	93.1	94.3	55.8	54.1	58.7	99.1	99.0	99.2
Penalized	0.962	0.958	0.966	90.5	89.4	91.5	93.6	93.0	94.1	56.1	54.0	58.0	99.1	99.0	99.2
Generalized additive model															
Degrees of freedom 4	0.963	0.958	0.967	90.8	89.6	92.0	93.4	93.0	94.1	55.3	53.9	58.0	99.1	99.0	99.2
Degrees of freedom 6	0.963	0.958	0.967	90.8	89.7	92.0	93.4	93.1	93.9	55.6	54.2	57.4	99.1	99.0	99.2
Degrees of freedom 8	0.963	0.958	0.967	90.8	89.5	92.0	93.5	93.1	94.5	55.7	54.3	59.4	99.1	99.0	99.2

Table 5C. Association measures and their 95% confidence intervals for claims-based algorithms derived from statistical methods with a condition-specific variable selection according to dyslipidemia

Method	AUC			Sensitivity			Specificity			PPV			NPV		
		95% CI		%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Regression															
Linear: model 1	0.709	0.705	0.714	49.9	49.2	50.8	90.2	89.9	90.6	76.4	75.7	77.2	73.9	73.7	74.3
Linear: model 2	0.739	0.734	0.744	43.7	42.6	45.1	91.2	90.0	92.1	75.9	74.1	77.7	71.9	71.5	72.2
Logistic: model 1	0.710	0.706	0.715	49.9	49.2	50.7	90.2	89.8	90.5	76.4	75.5	77.0	73.9	73.6	74.3
Logistic: model 2	0.738	0.733	0.743	43.6	42.4	44.9	91.3	90.3	92.2	76.0	74.3	77.7	71.8	71.5	72.1
Discriminant analysis															
Linear	0.739	0.734	0.744	43.6	42.5	44.9	91.2	89.9	92.2	75.9	74.0	77.8	71.8	71.5	72.2
Flexible	0.746	0.741	0.751	44.8	43.8	45.7	92.5	91.8	92.9	79.0	77.8	79.8	72.5	72.2	72.8
Penalized	0.739	0.734	0.744	43.7	42.3	45.1	91.3	90.0	92.2	76.1	74.0	77.7	71.9	71.5	72.2
Generalized additive model															
Degrees of freedom 4	0.758	0.753	0.763	49.4	48.6	50.3	90.7	90.4	91.1	77.2	76.5	77.9	73.9	73.6	74.2
Degrees of freedom 6	0.758	0.753	0.763	49.6	48.9	50.4	90.6	90.2	91.0	77.0	76.2	77.7	73.9	73.6	74.2
Degrees of freedom 8	0.758	0.753	0.763	49.6	48.7	50.5	90.6	89.9	91.3	76.9	75.9	78.2	73.9	73.6	74.2

Abbreviations: AUC, area under the receiver operating characteristic (ROC) curve; CI, confidence interval; NPV, negative predictive value; PPV, positive predictive value.

The number of observations of each of the diagnostic code and the medication code was used as input variables for model 1 in regression. Age and gender were added to these input variables for the other models.

Three types of discriminant analysis were conducted: linear discriminant analysis (LDA); flexible discriminant analysis (FDA); and penalized discriminant analysis (PDA). Multivariate adaptive regression splines with the second degree of interaction were used as the basis function in the FDA. In the PDA, a linear basis with the L_2-penalty was used, and the regularization coefficient was determined by cross-validation.

The additive logistic regression model with a cubic smoothing spline for each input variable except gender was used in the generalized additive model (GAM). The hyperparameter, the degrees of freedom for each smoothing spline, was set to the same value for all splines. I selected three values for the degrees of freedom, four, six, and eight, to generate three GAM prediction functions.

I calculated 95% CI for the AUC with Delong's method. 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 95% CIs for the resulting sensitivity, specificity, PPV, and NPV were calculated with 200 bootstrap resampling and the averaging methods.

Table 6A. Association measures and their 95% confidence intervals for claims-based algorithms derived from statistical methods without a condition-specific variable selection according to hypertension

Method	AUC			Sensitivity			Specificity			PPV			NPV		
		95% CI		%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Logistic regression															
Condition-specific	0.924	0.920	0.927	78.3	77.3	79.2	96.2	95.6	96.7	87.6	86.0	89.0	92.8	92.5	93.0
General	0.915	0.912	0.919	77.6	76.8	78.9	96.3	95.1	96.7	87.9	84.6	88.9	92.6	92.3	92.9
k-nearest neighbor															
Vote	0.915	0.911	0.919	77.5	76.1	78.4	95.7	94.9	96.5	86.3	84.0	88.4	92.5	92.1	92.8
Vote-Standardized	0.856	0.851	0.860	71.5	70.1	77.2	84.7	78.8	86.1	61.6	55.7	63.3	89.6	89.2	90.9
IDW	0.914	0.910	0.918	77.8	76.6	78.6	95.5	95.0	96.3	85.6	84.4	87.9	92.6	92.3	92.8
IDW-Standardized	0.855	0.850	0.859	72.9	70.3	76.2	83.2	79.3	85.6	59.9	56.1	62.7	89.9	89.3	90.7
Support vector machine															
Hinge loss	0.914	0.910	0.918	78.4	77.4	80.1	95.4	93.8	95.8	85.4	81.6	86.4	92.8	92.4	93.2
Squared hinge loss	0.919	0.915	0.923	79.5	78.7	80.4	95.8	95.2	96.0	86.6	85.1	87.3	93.1	92.9	93.4
Penalized regression															
Logistic Ridge	0.893	0.889	0.897	78.9	77.6	80.3	86.6	85.2	87.3	66.9	65.0	68.0	92.2	91.9	92.6
Logistic Lasso	0.924	0.920	0.927	78.6	77.6	79.5	96.1	95.2	96.5	87.3	85.0	88.6	92.9	92.6	93.1
Logistic Elastic-net	0.923	0.920	0.927	78.6	77.7	79.8	95.9	94.7	96.2	86.9	83.8	87.8	92.9	92.6	93.2
Tree-based model															
Random Forest-Breiman	0.923	0.920	0.927	80.8	79.9	81.7	95.5	95.0	96.0	86.1	84.8	87.3	93.5	93.2	93.8
Random Forest-Best AUC	0.923	0.919	0.927	80.7	79.8	81.8	95.8	94.8	96.2	86.7	84.4	88.0	93.5	93.2	93.8
ISLE-GBM learn 1	0.928	0.924	0.931	81.2	80.3	82.0	94.9	94.7	96.0	84.7	84.0	87.3	93.6	93.3	93.9
ISLE-GBM learn 0.05	0.930	0.927	0.934	81.5	80.7	82.2	95.0	94.7	95.4	85.0	84.2	85.9	93.7	93.5	93.9
ISLE-SGBM sample 0.5	0.930	0.927	0.933	81.3	80.3	82.2	95.0	94.6	95.7	85.0	83.8	86.5	93.6	93.4	93.9
ISLE-SGBM sample 0.1	0.929	0.926	0.933	81.0	80.1	81.9	95.3	94.9	95.8	85.6	84.5	87.0	93.6	93.3	93.8
Neural network															
Hidden units 5	0.912	0.908	0.916	78.0	76.9	78.7	95.5	95.1	95.8	85.5	84.5	86.6	92.6	92.3	92.9
Hidden units 10	0.914	0.911	0.918	78.6	77.8	79.4	95.4	95.0	96.1	85.6	84.6	87.2	92.8	92.6	93.1
Hidden units 20	0.910	0.906	0.914	78.4	77.6	79.2	94.8	94.5	95.3	84.0	83.1	85.2	92.7	92.5	93.0

Table 6B. Association measures and their 95% confidence intervals for claims-based algorithms derived from statistical methods without a condition-specific variable selection according to diabetes

Method	AUC			Sensitivity			Specificity			PPV			NPV		
		95% CI		%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Logistic regression															
Condition-specific	0.958	0.953	0.963	87.5	86.3	88.8	95.3	94.7	95.8	62.6	59.7	65.0	98.8	98.7	98.9
General	0.936	0.929	0.943	85.5	83.9	87.0	95.0	94.3	96.2	60.6	57.6	66.9	98.6	98.5	98.8
k-nearest neighbor															
Vote	0.942	0.936	0.947	83.9	82.0	85.8	95.8	93.5	96.6	64.1	54.2	68.9	98.5	98.3	98.7
Vote-Standardized	0.888	0.881	0.895	78.9	75.2	80.4	83.9	83.5	87.2	30.7	30.0	34.7	97.8	97.5	97.9
IDW	0.942	0.936	0.948	84.7	82.6	86.3	95.0	93.9	96.6	60.0	55.9	68.4	98.6	98.4	98.7
IDW-Standardized	0.889	0.882	0.896	78.3	76.7	82.2	85.3	81.7	85.9	32.3	28.7	33.2	97.8	97.6	98.1
Support vector machine															
Hinge loss	0.944	0.937	0.950	86.8	85.6	88.2	95.4	94.7	95.8	63.0	59.7	65.2	98.8	98.7	98.9
Squared hinge loss	0.950	0.944	0.956	88.2	86.8	89.5	95.1	94.0	95.7	61.9	57.2	64.7	98.9	98.8	99.0
Penalized regression															
Logistic Ridge	0.930	0.924	0.935	83.9	81.1	85.8	88.7	86.5	91.9	40.1	36.3	47.4	98.4	98.2	98.5
Logistic Lasso	0.961	0.956	0.965	88.7	87.4	89.8	94.9	94.4	95.3	61.0	58.7	62.8	98.9	98.8	99.0
Logistic Elastic-net	0.961	0.956	0.965	89.1	87.9	89.9	94.7	94.3	95.1	60.3	58.3	62.1	99.0	98.9	99.0
Tree-based model															
Random Forest-Breiman	0.960	0.956	0.965	88.5	86.8	89.7	95.0	94.4	96.0	61.6	58.7	66.4	98.9	98.8	99.0
Random Forest-Best AUC	0.958	0.954	0.963	88.6	87.3	89.8	95.0	93.4	95.4	61.2	55.2	63.3	98.9	98.8	99.0
ISLE-GBM learn 1	0.963	0.958	0.967	89.4	88.2	91.0	95.0	93.9	95.6	61.7	56.9	64.6	99.0	98.9	99.1
ISLE-GBM learn 0.05	0.965	0.961	0.970	90.3	89.0	91.3	94.1	93.7	95.1	58.2	56.2	62.5	99.1	99.0	99.2
ISLE-SGBM sample 0.5	0.965	0.961	0.969	90.5	88.8	91.7	93.8	93.4	95.3	57.0	55.2	63.0	99.1	99.0	99.2
ISLE-SGBM sample 0.1	0.963	0.959	0.968	89.9	88.6	90.8	94.2	93.9	95.4	58.5	57.0	63.4	99.0	98.9	99.1
Neural network															
Hidden units 5	0.919	0.912	0.926	80.9	79.6	82.4	95.4	94.5	96.1	61.3	57.0	65.1	98.2	98.1	98.4
Hidden units 10	0.939	0.933	0.945	82.9	81.3	85.0	95.4	93.9	96.5	62.1	55.4	68.0	98.4	98.3	98.6
Hidden units 20	0.934	0.927	0.940	84.0	82.3	85.8	95.5	93.7	96.6	62.5	54.8	68.8	98.5	98.4	98.7

Table 6C. Association measures and their 95% confidence intervals for claims-based algorithms derived from statistical methods without a condition-specific variable selection according to dyslipidemia

Method	AUC			Sensitivity			Specificity			PPV			NPV		
		95% CI		%	95% CI		%	95% CI		%	95% CI		%	95% CI	
Logistic regression															
Condition-specific	0.738	0.733	0.743	43.6	42.4	44.9	91.3	90.3	92.2	76.0	74.3	77.7	71.8	71.5	72.1
General	0.743	0.738	0.748	49.9	46.4	55.8	85.2	79.5	88.4	68.1	63.1	72.0	72.8	72.2	73.9
k-nearest neighbor															
Vote	0.739	0.734	0.744	48.3	45.1	51.9	87.3	83.7	90.8	70.8	67.1	75.4	72.7	72.2	73.3
Vote-Standardized	0.680	0.674	0.685	50.3	46.6	54.3	74.3	70.2	77.8	55.5	53.6	57.4	70.2	69.5	70.8
IDW	0.739	0.733	0.744	48.7	46.1	51.6	87.2	84.3	89.6	70.6	67.5	73.7	72.8	72.3	73.4
IDW-Standardized	0.677	0.672	0.683	50.1	46.4	53.5	74.6	71.3	78.1	55.6	53.8	57.6	70.2	69.5	70.9
Support vector machine															
Hinge loss	0.724	0.719	0.730	49.8	49.1	50.7	90.3	89.9	90.8	76.6	75.7	77.4	74.0	73.7	74.3
Squared hinge loss	0.749	0.744	0.754	51.2	47.9	56.1	85.2	80.2	87.9	68.6	64.2	71.7	73.3	72.6	74.3
Penalized regression															
Logistic Ridge	0.725	0.719	0.730	56.1	51.4	66.2	75.3	65.1	80.0	59.2	54.6	62.2	73.0	72.0	75.2
Logistic Lasso	0.753	0.748	0.758	49.6	48.6	50.4	90.6	90.0	91.0	76.9	76.0	77.8	73.9	73.5	74.2
Logistic Elastic-net	0.748	0.743	0.753	48.6	44.3	53.3	87.0	81.8	91.6	70.5	65.4	76.8	72.7	72.0	73.5
Tree-based model															
Random Forest-Breiman	0.761	0.756	0.766	50.1	48.9	52.0	90.4	88.8	91.4	76.8	74.3	78.4	74.1	73.7	74.5
Random Forest-Best AUC	0.760	0.755	0.765	50.1	49.1	51.3	90.5	89.2	91.0	77.0	75.0	78.0	74.0	73.7	74.4
ISLE-GBM learn 1	0.767	0.762	0.772	50.7	49.0	52.7	89.4	87.6	90.8	75.3	72.8	77.2	74.1	73.7	74.5
ISLE-GBM learn 0.05	0.772	0.767	0.777	50.1	48.8	52.3	90.6	88.7	91.4	77.2	74.3	78.5	74.1	73.7	74.6
ISLE-SGBM sample 0.5	0.772	0.767	0.777	50.8	49.0	53.6	89.8	87.3	91.7	76.0	72.6	78.8	74.2	73.8	74.9
ISLE-SGBM sample 0.1	0.768	0.763	0.773	50.4	48.3	52.8	89.8	87.7	91.7	75.9	73.0	78.8	74.0	73.6	74.6
Neural network															
Hidden units 5	0.739	0.734	0.744	54.5	47.0	59.6	80.7	75.7	87.8	64.2	60.7	71.1	73.7	72.3	74.5
Hidden units 10	0.745	0.739	0.750	49.5	47.0	51.5	87.5	85.8	89.8	71.5	69.4	74.8	73.2	72.7	73.6
Hidden units 20	0.741	0.736	0.747	50.4	44.1	56.1	85.5	79.8	91.5	68.8	63.7	76.6	73.1	72.1	74.1

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

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 condition-specific in the logistic regression. Condition-specific in the logistic regression is the same as model 2 of the logistic regression. General in the logistic regression fitted a logistic regression model to the dataset that was trimmed properly.

The Euclidean distance with raw or standardized (i.e., re-scaled 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), the L_1-penalty (logistic lasso), and the elastic-net penalty (logistic elastic-net) were applied. The regularization coefficient and the elastic-net mixing parameter were determined by cross-validation.

Two types of tree-based models were applied: the random forest and the importance sampled learning ensemble (ISLE). The minimum node size was set to ten for each tree, and two-hundred trees were bagged in the random forest. The number of variables selected for each split was first set to the value recommended by Breiman, $\lfloor \sqrt{p} \rfloor$, where p is the number of input variables, and then tuned using the validation set next. Denote the ISLE with the subsampling ratio equals to one and less than one by ISLE-gradient boosting machine (GBM) and ISLE-stochastic gradient boosting machine (SGBM), respectively. I fixed the depth to be six for all ISLEs. For the ISLE-GBM, I selected the learning rate to be 1 and 0.05. For the ISLE-SGBM, I fixed the learning rate to be 0.1 and selected the subsampling ratio to be 0.5 and 0.1. The number of trees and the 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: five, ten, and twenty. All hidden units were fully connected with the nodes in the input and output layer. Weight decay was employed for the regularization of the parameters, and the regularization coefficient of it was tuned using the validation set.

I calculated 95% CI for the AUC with Delong's method. 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 95% CIs for the resulting sensitivity, specificity, PPV, and NPV were calculated with 200 bootstrap resampling and the averaging methods.