論文の内容の要旨

論文題目　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　原　湖楠

1 Introduction

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. Nevertheless, claims data is subject to limitations due to potential imprecision in the identification of medical conditions. 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. 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.

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.

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 (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.

2 Methods

2.1 Setting

Medical and pharmacy claims data combined with annual health screening results were obtained from Japan Medical Data Center. 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, (2) with complete data on self-reported use of blood pressure-lowering drugs, hypoglycemic drugs, and lipid-lowering drugs for FY2016 and FY2017, (3) who in FY2017 visited a clinic/hospital that mainly specializes in internal medicine, 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). I constructed a gold standard from the results of the annual health screening. 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.

2.2 Claims-based algorithm

The CBA was compared with the gold standard. 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.

Conventionally, researchers have selected input variables and decided how to incorporate variables into the CBA by hand. Hence, I first developed three conventional case-finding algorithms for each condition as baseline CBAs. 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 and C09, A10, and C10. To evaluate to what extent baseline CBAs is applicable to a wide range of populations, the following study populations were considered instead of the baseline study population: (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.

Statistical methods such as regression and statistical learning methods can foster the development of CBAs. Accordingly, I next 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. Consequently, I lastly 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.

2.3 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). 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. For the CBAs that were based on statistical methods, a prediction function needs to be derived 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 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. 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. All statistical analysis was conducted using R version 3.5.1. R code will be available at https://github.com/harakonan/research-public/tree/master/cba after the publication of the study.

3 Results

As the test set was employed for the calculation, the sample size was 157,822, 38,092, and 153,608 for hypertension, diabetes, and dyslipidemia. The prevalence that was determined by the gold standard for each condition was 25.4%, 8.3%, and 38.7% for hypertension, diabetes, and dyslipidemia.

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. The sensitivity decreased when the study population was expanded to include all people (hypertension, 66%-71%; diabetes, 74%-85%; dyslipidemia, 28%-38%).

The AUC of the regression model (discriminant analysis and GAM) with a dataset that input variables were selected according to each condition was .924-.925 (.925-.929) for hypertension, .958-.962 (.962-.963) for diabetes, and .738-.739 (.739-.758) for dyslipidemia.

The AUC of the models with a dataset that input variables were chosen to be common to all target conditions was as follows: the logistic regression, hypertension .915, diabetes .936, dyslipidemia .743; the kNN with raw (standardized) input variables, .914-.915 (.855-.856), .942 (.888-.889), .739 (.677-.680); the SVM, .914-.919, .944-.950, .724-.749; the logistic ridge (the logistic lasso and the logistic elastic-net), .893 (.923-.924), .930 (.961), .725 (.748-.753); the random forest (the ISLE), .923 (.928-.930), .958-.960 (.963-.965), .760-.761 (.767-.772); the neural network, .910-.914, .919-.939, .739-.745.

4 Discussion

As I expanded the study population to include all enrollees from the baseline study population, the sensitivity decreased. 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%).

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.

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. 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. 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.

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.

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.