Early Detection of Diabetes from Health Claims

Rahul G. Krishnan, Narges Razavian, Youngduck Choi
New York University
{rahul,razavian,yc1104}@cs.nyu.edu

Somesh Nigam
Independence Blue Cross
somesh.nigam@ibx.com

Saul Blecker, Ann Marie Schmidt
NYU School of Medicine
{saul.blecker,annmarie.schmidt}@nyumc.org

David Sontag
New York University
dsontag@cs.nyu.edu

Abstract

Early detection of Type 2 diabetes poses challenges to both the machine learning and medical communities. Current clinical practices focus on narrow patient-specific courses of action whereas electronic health records and insurance claims data give us the ability to generalize that knowledge across large sets of populations. Advances in population health care have the potential to improve the quality of health of the patient as well as decrease future medical costs, at least in part by prevention of long-term complications accruing during undiagnosed diabetes. Based on patient data from insurance claims, we present the results of our initial experiments into identification of patients who will develop diabetes. We motivate future work in this area by considering the need to develop machine learning algorithms that can effectively deal with the depth and the variety of the data.

1 Introduction

Diabetes is an increasingly prevalent disease that affects millions of people in the United States and around the world. With burgeoning health care costs both to the individuals and to health care providers, there is an urgent need to develop efficient ways to detect diabetes and diabetes vulnerability before the frank diagnosis of this disorder. This allows providers to screen patients and advise remedial courses of action, ideally leading to prevention not only of diabetes but also the complications of diabetes including cardiovascular disease, kidney disease, peripheral vascular disease, and retinopathy.

Current clinical procedures have focused on evidence-based screening, where patients exhibiting risk factors for diabetes such as hypertension and obesity are screened. This approach has the distinct disadvantage of diagnosing patients at the individual level but not at the population level. As a result, many patients are often detected late in the progression of the disease, when preventive measures are significantly less effective.

The focus of this extended abstract is on the early detection of diabetes from health insurance claims. Our data are composed of a large cohort of 5 million individuals of age above 18 years. Patient records include medical, pharmacy and lab information related to the individuals between years 2005 and 2013. The large number of patients, and the availability of thousands of features makes the problem both challenging and interesting. Larger datasets allow us to perform experiments on a variety of experimental settings and subsets of the data without losing the significance of our conclusions. Performing large scale training and validation over noisy and incomplete data, on the other hand, is very challenging. In our work, we perform prediction using varying spans of patient history and varying windows into the future. This allows us to categorize the quality of the estimate as well as the effect of having more information to make the prediction.
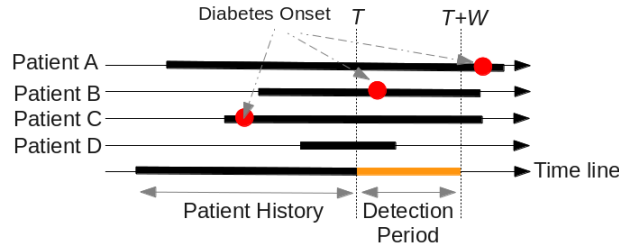
Figure 1: At time T we predict for patients not currently diabetic whether they will be diabetic at $T + W$. Bold shows the months that the patient is enrolled, and red dots denote the time of first diabetes diagnosis.

1.1 Related Work

Prediction of Type 2 diabetes has received significant attention in the research community. In a recent study, Abbasi et al. [1] performed a systematic literature survey and independent validation of over 25 prediction models, 12 of which use basic features (demographics, family history of diabetes, measures of obesity, diet, lifestyle factors, blood pressure and use of antihypertensive drugs), and the other 13 additionally including bio-markers such as HbA1c. As is common in clinical research, prediction models were learned using logistic regression. The area under the receiver operator characteristic curve (AUC) for predicting diabetes onset 7.5 years after baseline was found to be 0.84 for the best model in the first group, and 0.93 for the best model in the second group. However, none of the models could sufficiently quantify the actual risk of future diabetes (i.e., they were not well-calibrated).

In a recent online competition, Kaggle hosted a diabetes classification task using a dataset of 9948 members from Practice Fusion, a web-based electronic health record. The winning teams used boosted regression trees on a large set of clinical and derived features, including many of those described above. Our study differs both in the problem setup and in the type of data used. First, the Kaggle task was to identify members *already* diagnosed with type 2 diabetes, where diagnosis codes for diabetes, glucose lab tests, and diabetes medications are artificially removed from the data. However, it may be difficult to completely mask the diabetes diagnosis and treatment, making the prediction task ill-posed. Second, data found in electronic health records can differ from that found in health insurance claims. Electronic health records can often be incomplete since they are compiled by individual medical institutions. A patient may visit multiple clinics and hospitals over the course of an illness, and seldom would all of their medically relevant transactions be captured by a single institution. Since insurance companies pay for all medical transactions, their records can be more comprehensive and rigorously maintained.

Most closely related to our work is an earlier study by Neuvirth et al. [6], who also use claims data to study diabetes and show that it is possible to identify high-risk patients such as those likely to need emergency care services. We use a similar set of features as theirs for our initial algorithms.

2 Methodology

We are interested in understanding the effects of patient history, and how far into the future we can predict the onset of type 2 diabetes. We formulated this question as a series of experiments where we used varying patient history and a varying prediction window into the future. Figure 1 shows how we defined each history and prediction window. For each T and W pair, we trained a separate model to predict the probability of patients developing diabetes between time T and $T + W$. Patients who had developed diabetes before time T are excluded. For instance, patient C in the figure matches this criteria and is excluded for this T and W pair. A patient is given a positive diabetes label if they developed diabetes within the T and $T + W$ window. For instance in Figure 1, patient A is assigned a negative label, and patient B is assigned a positive label. We also excluded any patient who was continuously enrolled less than two-thirds of time period between T and $T + W$. As an example, patient D would be excluded from the evaluation of predictive accuracy for this T and W pair.

Feature	All (T=2010)	Diabetic (T=2010)	All (T=2008)	Diabetic (T=2008)
Average Age (std)	37.51 (19.87)	52.56 (15.55)	40.37 (20.23)	56.29 (15.61)
Gender	53% Female	55% Female	53% Female	51% Female
High BP	18.97%	44.82%	17.29%	41.44%
High Cholesterol	12.26%	25.63%	10.26%	24.07%
Obesity	5.36%	12.57%	4.42%	10.38%

Table 1: Characteristic Table for the dataset. The pairs of columns represent the statistics present in patients' full history up to the specified year. Diabetic patients include those diagnosed within two years of the specified year (i.e $W = 2$ years). Blood pressure, cholesterol and obesity features are based on IDC9 codes diagnosed by the physician and recorded in medical claim records.

Diabetes onset in the interval between T and $T + W$ was determined based on criteria developed by the clinical co-authors and modeled on previous research studies. We used $A1C$ levels, ICD9 codes and NDC codes prescribed for diabetes to determine the conditions to mark the onset of diabetes.

Traditional approaches for prediction of diabetes rely heavily on features such as body mass index (BMI), ethnicity, height and weight. These were four distinct features that are not cataloged in health insurance claims data. However, our dataset covers all health care encounters such as medical claims, lab results and drug prescriptions with specific details on lab values and dosages of prescriptions. Therefore, we remain optimistic that this rich source of information will give us the opportunity to explore and infer some of the less well known indicators of diabetes.

We defined two sets of features from patient history. The first consists of 22 features commonly used in the literature to predict diabetes [1, 2, 3, 8]. These include basic demographics, history of diagnosis of conditions such as high blood pressure or high cholesterol, and measurement of related bio-markers including fasting plasma glucose level, triglyceride level, and cholesterol in HDL. The second group of 1054 features were created based on a much larger set of data. In addition to the first set of features, we included medication history grouped according to both the standard and specific therapeutic class of medications (999 features), and the diagnosis history grouped according to the top-level of the ICD-9 hierarchy (34 features).

All the information about a patient was first aggregated into a single file where each line represented a patient. The feature vector for some value of T was created by considering the entire history of the patient up to time T. Many features are binary with 1 indicating at least one instance of the patient being associated with the ICD-9 code or the class of medication specified earlier. Non binary features included profile information such as age, number of medical records and number of drug prescriptions among others.

3 Experiments

Our entire dataset consists of more than 5 million patients, all above 18 years of age. This study was determined to be IRB exempt (IRB 13-9401: Machine Learning to Predict Undiagnosed Diabetes from Insurance Claims) since the data was de-identified and stripped of personal information. For the experiments presented herein, we used $330,000$ randomly selected patients from our database. Training was performed using regularized logistic regression, as implemented in the scikit-learn [7] package. We use a pre-determined train and test split for the experiments. We tune hyper-parameters using three-fold cross-validation to select the optimal regularization parameter and regularization type (L1 and L2 penalty functions). We create 95% confidence intervals for AUC values based on variances calculated using the nonparametric approach of Delong et al. [4, 9].

Table 1 presents the characteristics table of the data. The last three rows in the table are features that have been previously shown to be important for diagnosing diabetes. More specifically the All (2008) column corresponds to the entire dataset at $T = 2008$ while the Diabetic (2008) column corresponds to the subset of the dataset comprising individuals who were diagnosed with diabetes in the two year period after 2008. As we travel from left to right from All to Diabetic across columns we see that the percentages in the last three rows increase. This has a natural explanation since individuals with diabetes are well-known to be more likely to exhibit higher blood pressure, cholesterol and obesity.

T	W	AUC	AUC (age <45)	AUC (age >45)	# of Features
2008	3	0.807 ± 0.051	0.767 ± 0.115	0.743 ± 0.073	22
2008	3	0.798 ± 0.051	0.776 ± 0.113	0.727 ± 0.075	1054
2008	6	0.790 ± 0.040	0.766 ± 0.090	0.700 ± 0.060	22
2008	6	0.798 ± 0.039	0.777 ± 0.087	0.711 ± 0.058	1054
2008	12	0.776 ± 0.030	0.730 ± 0.075	0.672 ± 0.044	22
2008	12	0.782 ± 0.030	0.743 ± 0.074	0.682 ± 0.044	1054
2008	24	0.768 ± 0.025	0.742 ± 0.059	0.656 ± 0.036	22
2008	24	0.773 ± 0.024	0.751 ± 0.058	0.666 ± 0.035	1054
2010	3	0.781 ± 0.060	0.764 ± 0.136	0.688 ± 0.086	22
2010	3	0.763 ± 0.065	0.743 ± 0.138	0.678 ± 0.093	1054
2010	6	0.760 ± 0.047	0.727 ± 0.105	0.673 ± 0.064	22
2010	6	0.766 ± 0.046	0.738 ± 0.098	0.677 ± 0.065	1054
2010	12	0.752 ± 0.036	0.726 ± 0.074	0.651 ± 0.052	22
2010	12	0.756 ± 0.035	0.743 ± 0.069	0.655 ± 0.051	1054
2010	24	0.747 ± 0.030	0.716 ± 0.060	0.640 ± 0.043	22
2010	24	0.750 ± 0.029	0.723 ± 0.058	0.647 ± 0.042	1054

Table 2: Comparison of AUC (with 95% CI) evaluated on test data for different starting times (T) and prediction windows W (in months). See Fig. 1 for a visual depiction of T and W in the experiments.

Table 2 shows the area under the ROC curve (AUC) for different lengths of patient history T, and prediction windows W (in months). To further analyze the results, we computed the AUC values over different subsets of the data. Stratifying based on age allows us to consider the quality of our predictor on a young population versus an older population. While this method is not a rigorous analysis of age-matched populations with and without diabetes, it is indicative that that an older population represents a harder prediction task for the algorithm, as age could no longer be used to quickly disambiguate healthy from potentially sick patients. This accounts for the decrease in AUC shown in Table 2 as we move from the slice of data representing young people to that representing older people. The table also compares the results obtained from classification with 1054 features versus 22 features. As expected, we see that most of the AUC values increase with the use of a larger set of features, however this increase is not as marked as we might have hoped. This suggests that there still exists room for significant improvement in the algorithm by improved feature engineering and principled approaches to dealing with problems such as missing and noisy data.

4 Discussion

Our initial experiments and baseline statistics have helped us establish several ideas. The first is that looking at patients across values of T and W allow us to analyze changes in the distribution over people and the difficulty of predicting far into the future, while also keeping the prediction task realistic. The second is that the reduced set of features in insurance claims relative to traditional clinical data has prompted us to look for informative features in other locations. This problem also motivates the need to find structure in and analyze situations where we have large quantities of noisy and missing data, which begs the question of whether one can automatically infer these quantities. We hope to expand on prior work [5] done in this area. In particular, we plan to consider a temporal class of models that potentially could be used to infer some hidden state about the patient.

One shortcoming of our current evaluation methodology is that the predictor for time T is learned using data in the range from T to $T + W$ (i.e., peeking into the future) to obtain the diabetes labels for the training points. Instead, we should be using a model trained solely using retrospective data. As we illustrated in Table 1, this may require us to correct for changes in the underlying distribution.

In conclusion, we have early results that indicate that health insurance claims data is a rich source of information for the early detection of diabetes. One set of particularly interesting questions has to do with obtaining a better understanding of the disease. For example, why is it that some diabetic patients go on to develop significant cardiovascular disease, whereas others do not? Could we use claims data to obtain new insights into the disease mechanism that could lead to new treatments? Many other clinical questions can also be studied using this data, notably discovering how chronic illnesses such as diabetes, kidney disease, or congestive heart failure evolve over time.

References

[1] Ali Abbasi, Linda M Peelen, Eva Corpeleijn, Yvonne T van der Schouw, Ronald P Stolk, Annemieke M W Spijkerman, Daphne L van der A, Karel G M Moons, Gerjan Navis, Stephan J L Bakker, and Joline W J Beulens. Prediction models for risk of developing type 2 diabetes: systematic literature search and independent external validation study. *BMJ*, 345, 9 2012.

[2] Muhammad A Abdul-Ghani, Ken Williams, Ralph A DeFronzo, and Michael Stern. What is the best predictor of future type 2 diabetes? *Diabetes care*, 30(6):1544–1548, 2007.

[3] Gary S Collins, Susan Mallett, Omar Omar, and Ly-Mee Yu. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine*, 9(1):103, 2011.

[4] DeLong ER, DeLong DM, and Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operativng characterstic curves: a non parametric approach. *Biometrics*, 44:837–845, 1988.

[5] Yoni Halpern and David Sontag. Unsupervised learning of noisy-or bayesian networks. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence (UAI-13)*. UAI Press, 2013.

[6] Hani Neuvirth, Michal Ozery-Flato, Jianying Hu, Jonathan Laserson, Martin S Kohn, Shahram Ebadollahi, and Michal Rosen-Zvi. Toward personalized care management of patients at risk: the diabetes case study. In *Proceedings of the KDD*, pages 395–403, 2011.

[7] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[8] Peter WF Wilson, James B Meigs, Lisa Sullivan, Caroline S Fox, David M Nathan, and Ralph B D'Agostino Sr. Prediction of incident diabetes mellitus in middle-aged adults: the framingham offspring study. *Archives of Internal Medicine*, 167(10):1068, 2007.

[9] Robin X, Turck N, and Hainard A. pROC: an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, pages 12–77, 2011.