Methods for Predicting Type 2 Diabetes

CS229 Final Project
December 2015

Duyun Chen[1], Yaxuan Yang [2], and Junrui Zhang [3]

Abstract

Diabetes Mellitus type 2 (T2DM) is the most common form of diabetes [WHO (2008)]. More than 29 million people in the United States are affected by T2DM and another 86 million are in a state of prediabetes, a condition that exhibits high risk to progress into diabetes [NIH (2014)]. Many T2DM cases can be prevented or avoided by improved awareness and lifestyle adjustments [NIH (2014)]. Our project aims to improve T2DM diagnosis methodologies using supervised machine learning algorithms trained on data from electronic medical records (EMR). Specifically, SVM, Adaptive Boosting with Decision Trees, Random Forests, and Logistic Regression were used to build models for predicting T2DM.

Keywords

Diabetes, Machine Learning, Logistic Regression, SVM, Random Forest, Decision Tree, AdaBoost

Introduction

More than 29 million people in the United States are affected by type 2 Diabetes Mellitus (T2DM), and many cases can be prevented or avoided by improved awareness and lifestyle adjustments. The goal of this project is to build a model for predicting T2DM. Kaggle offers a vast feature set ranging from medical history to lifestyle data provided by Practice Fusion, one of the fastest growing Electronic Health Record communities.

After scrutinizing the raw data set, three main challenges were identified:

1. Feature selection - The data exhibits high diversity, ranging from physical attributes and demographics to personal data such as smoking history and medication profile. Therefore, feature selection and engineering were very important. We chose to first select features based on domain knowledge (all team members hold degrees in biological fields) and literature review. In a second pass, the features were further pruned using both bottom-up and top-down sequential feature selection methods.
2. Data imbalance and invalidity - Because only 19% of the patients in the dataset are designated DMT2 positive, the data set is imbalanced and will affect the final training model. To address this problem, both undersampling and oversampling techniques were used to rebalance the data set. F1 score is used to evaluate the model performance. Moreover, approximately half of the records exhibit incorrect information such as negative blood pressure and impossibly low height/weight. Because the unuseable data made up a significant fraction of the data set, the focus was primarily on the raw data set, but Logistical Regression was used to analyze the cleaned data set for comparison.
3. Nonlinear classification - Principle Component Analysis (PCA) of the preliminary selected features with 2 components revealed no clear decision boundary between the classes, suggesting linear classifiers will not be an optimal choice. Therefore, instead of arbitrarily selecting classification methods, several specific classification models that generally work well with such issues were chosen: Support Vector Machine (with various Kernels), Random Forest, and Adaptive Boosting (AdaBoost) with Decision Tree as the weak learner.

Data and Methods

Dataset, Preliminary Feature Extraction and Feature Engineering

This project used a publicly available EMR dataset released by Practice Fusion in 2012 for a Kaggle competition [Kaggle (2012)]. It consists of electronic health records for 9,948 patients, among whom 1,904 have been diagnosed with DMT2.

First, a relational database was constructed by extracting data from 6 tables in the Kaggle dataset. 10 raw features were selected as preliminary inputs for model training, including BMI, Gender, Age, SystolicBP (systolic blood pressure), DiastolicBP (Diastolic blood pressure), NdcCode (National Drug Directory medication code), Smoking status, ICD9code (International Statistical Classification of Diseases and Related Health Problems, Diagnosis Code), HL7 (Health Level 7) Identifier and isAbnormalValue (abnormal lab result). These features were chosen based on domain knowledge obtained through literature review. To note, BMI is a derived variable from two of the original features, Height and Weight.

Different approaches have been taken to engineer categorical features. For NdcCode, binary features were used

[1]Department of Computer Science, Stanford University
[2]Department of Statistics, Stanford University
[3]Department of Computer Science, Stanford University
Email: duchen@stanford.edu
yangyax@stanford.edu
junrui@stanford.edu

to indicate if a patient has taken a specific medication. To reduce the dimensionality, only the top 10 most used medications were considered. For ICD9code, 19 binary features were chosen to indicate if a patient has been diagnosed with a certain disease belonging to these 19 classes. These related diseases were chosen based on literature. For HL7Identifier (blood tests), all labs listed in the data set were screened by researching the relevance to DMT2. 12 lab tests were selected because of supporting correlations between abnormal concentrations of these compounds in the blood stream and DMT2. [Eschwege, Richard, Thibult (1985)]

Through this preliminary filter using domain knowledge and literature support, 46 features out of over 500 were selected for model construction and further analysis.

Feature Normalization and Cleanup

In the provided training dataset, several features have missing values. These cases were handled by setting them to 0.

By examining the distributions of all features, a few abnormalities have been identified (Figure 1). For example, despite the patients all being adults, approximately half of them are shorter than 40 inches in height. Also, over 10% of the patients weigh less than 20 pounds. In addition, negative blood pressures were observed in some cases. Since eliminating these records dramatically reduced the data set (approximately 50%), only logistically regression was trained on the cleaned data set for a simple comparison.

Dataset Partition and Rebalancing

20% of the original dataset was used as validation set for feature selection, and 20% was taken as a test set for generalization error evaluation. The remaining 60% was used as the training set.

Two different techniques were chosen to rebalance the class labels: undersampling and oversampling. Undersampling was performed by Random majority undersampling with replacement, and the ratio of two classes was balanced to 1.0 by pruning the data. For oversampling, Synthetic Minority Oversampling Technique (SMOTE) was used to balance the ratio of the two classes to 0.998. The *UnbalancedDataset* package (*https://github.com/glemaitre/UnbalancedDataset*) was used to perform the balancing. The undersampling technique reduced the patients without diabetes from 4868 to 1125, while the oversampling method increased the diabetic samples from 1125 to 4861.

Models

A logistic regression classifier was trained using the *glm* package in R. Additional data quality check was performed to remove extreme outliers and/or abnormal values (see). Lasso regularization was performed to auto select features.

The *SciKit* package was used to do the SVM, Random Forest and AdaBoost analyses (*http://scikit-learn.org/*). For SVM, all kernels supported by this package were used in the analysis: linear, polynomial (degree 3), sigmoid, and RBF (Gaussian, radial basis function). Also, the regularization parameter was varied between 0.01 and 10.0. For Random

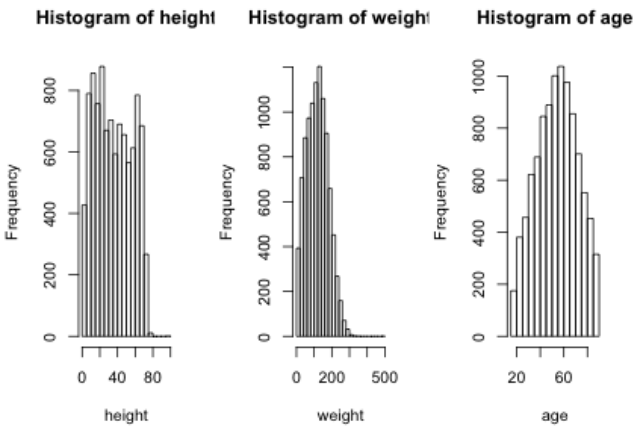


Figure 1. Distributions of three numerical features in the raw data set: height, weight and age. Note that approximately half of the patients exhibit abnormal height records: < 40 inches despite being all adults ($\geq$ 18 yr old). Also, there are roughly 1100 patients who weigh under 20 lbs. These are considered incorrect observations and will be excluded in logistic regression analysis. They were not excluded from other models due to how much the data would be reduced.

Forest, 25 trees were used. For AdaBoost, Decision Tree was chosen as the weak learner, and the SAMME discrete boosting algorithm was used. As default settings, the maximal number of estimators at termination is 200, and maximal Decision Tree height was set to 1. For later parameter adjustment with AdaBoost, the maximal tree height was varied from 1 to 10, and the number of estimators was tested from 10 to 700 with a learning rate of 30.

Principle Component Analysis

PCA was performed using *SciKit* package (*http://scikit-learn.org/stable/*), and two major components were selected for visualization.

Results and Discussion

Model Selection and Data Balancing

To get a preliminary understanding of the targeted problem, we performed PCA visualization using the top 2 major components (figure not shown), and no possible decision boundary was observed. Therefore, we decided to only use discriminative non-linear classifiers such as SVM, Random Forest, AdaBoost. In addition, logistic regression was used to check the effect of data cleanup.

The dataset was balanced using both undersampling and oversampling. We found undersampling worked better than the raw skewed data. For example, using AdaBoost, skewed data has a F1 score of 0.3125 while undersampled data has a F1 score of 0.4637. Oversampling generally worked better than undersampling, so oversampled dataset was used by default for training and validation.

Learning Curve Analysis

As most of the features were selected using domain knowledge, some features could be redundant, which can lead to overfitting. Thus, a learning curve analysis using different sized training dataset (balanced by oversampling)

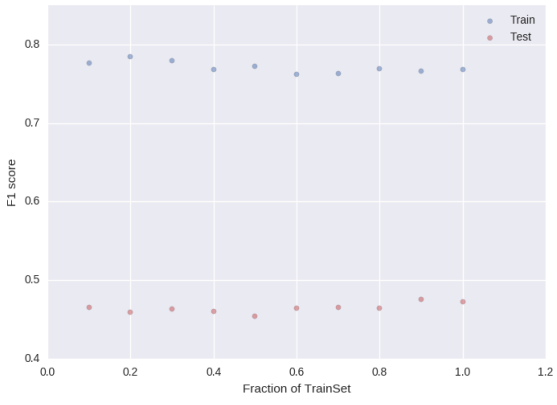


Figure 2. Learning Curves of the AdaBoost learner on training sets of different sizes. The curves are evaluated by F1 Score. Blue: training data; Red: test data.

with an AdaBoost learner was implemented and the results are shown in Figure 2. The learning curve concluded the size of the training dataset did not significantly affect the generalized F1 scores, suggesting that high-variance is not an issue. Therefore, retaining redundant features is not a problem. Similar results were also observed for the learning curve analyses on logistic regression and SVM models. Together, this result suggests a bias issue rather than a variance issue.

However, since precision and recall was used instead of other evaluation methods to plot the learning curves, the results here may not completely rule out the existence of redundant features. Therefore, feature selection was done despite the above results.

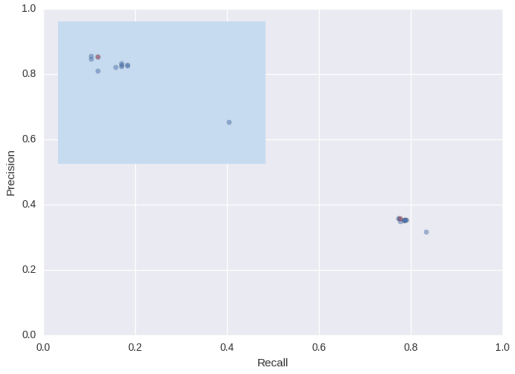
Feature Selection

Both bottom-up and top-down sequential feature selection algorithms were tested on SVM, Random Forest and AdaBoost learners. We were also interested in how much the undersampling and oversampling techniques improved the learners' performances, so they were evaluated simultaneously. The results of the Random Forest and AdaBoost models are shown as Precision-Recall curves (PR curves) in Figure 3 and Table 2. We found different learners had different weights on features.

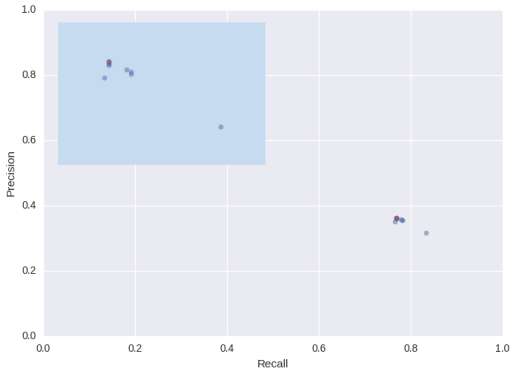
Judging from the p-value of the coefficients, the most significant features of the logistic regression are Height, BMI, Weight, Diastolic/Systolic Blood Pressures, Age, Gender, abnormal triglyceride level, and previous diagnosis of metabolic/immune/circulatory diseases.

The top 5 features that made the most significant contributions to the SVM model are (in order of significance): Gender, Weight, Height, Age, Systolic Blood Pressure. These are not surprising, as the literature supports them as factors for T2DM. They were computed using top-down feature selection and choosing the highest decreases in F1 after removing certain features.

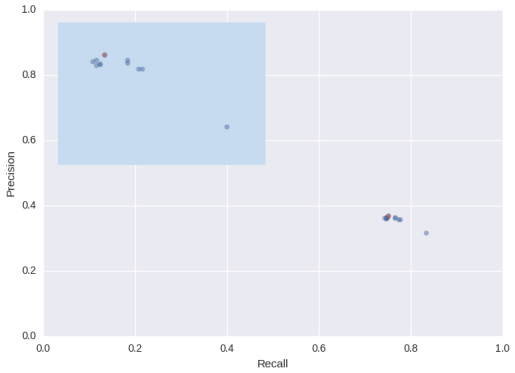
The top 5 features for the Random Forest model are: circulatory diseases, metabolic diseases, blood chloride level, albuterol treatment and SeptraDS treatment. It is not surprising that circulatory and metabolic diseases are the most important indicators as diabetic patients have an



(a) RF Bottom-Up Undersampling



(b) RF Bottom-Up Oversampling

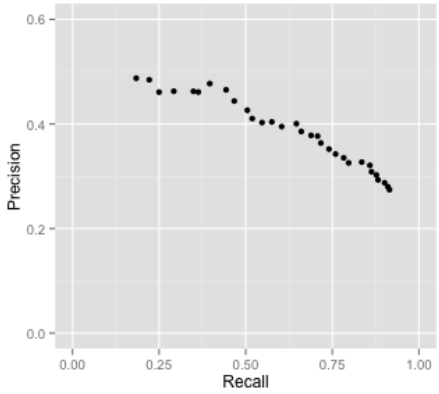


(c) Ada Bottom-Up Oversampling

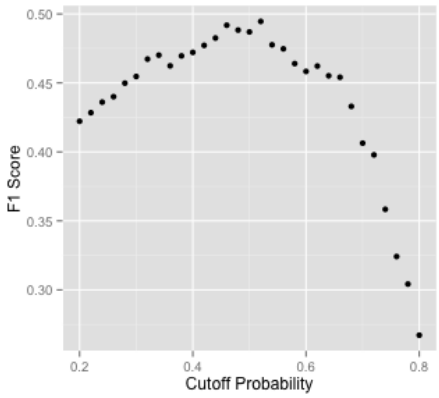
Figure 3. Precision-Recall curves of feature selections. The inlet graph is the zoom-in of the pictures. The red dot represents the selected feature set. RF: random Forest; Ada: AdaBoost; Top-Down: Top-Down sequential feature selection.

extremely high chance to also suffer from life-threatening circulatory diseases (84% of people 65+ years old) [AHA (2014)], while diabetes itself is a metabolic disease. Albuterol can induce high blood sugar levels [ProAir (2010)] while there is no reported correlation between diabetes and SeptraDS treatment [Mihic,Mautner,Feness,Grant (1975)] or chloride levels.

The top 5 features for the AdaBoost model are: circulatory diseases, metabolic diseases, blood Potassium level, genitourinary/urinary diseases and ill-defined disease conditions. Potassium plays a very important role in diabetes [Chatterjee,Yeh,Edelman,Brancati (2011)]. It is also well-known that diabetes will stress the urinary system, and can lead to many ill-defined symptoms.

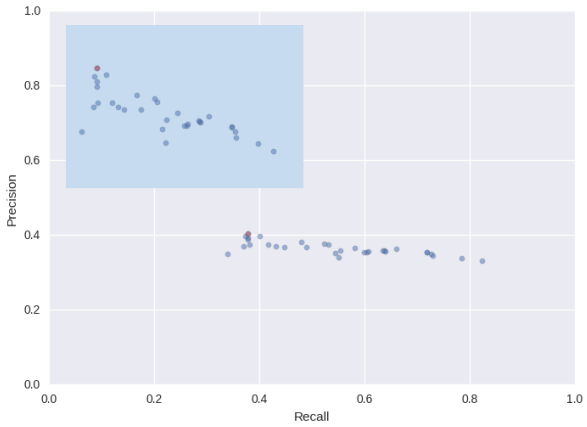


(a) Precision-Recall Curve with Respect to Various Cutoff Probabilities

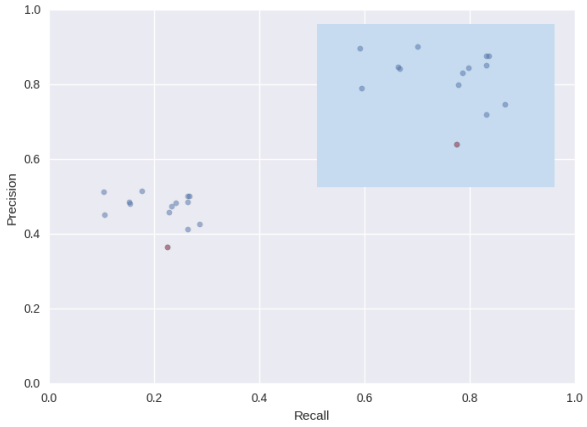


(b) F1 scores with respect to different cutoff probabilities

Figure 4. Precision, recall and F1 score of the logistic regression classifier with respect to different cutoff probabilities. (a) Precision-Recall value evaluated at a range of different cutoff probabilities. (b) Relationship between F1 score and cutoff probability



(a) Over-Sampled Data



(b) Raw Data

Figure 5. Precision-Recall curve of parameter search with AdaBoost with Decision Tree on oversampled (a) and raw dataset (b). Red: selected point.

Consistent with the varied feature weights across different models, the feature selection was neither very effective nor readily repeatable (Figure 3), which also agrees with the result observed in the learning curve: the learners did not exhibit a high variance issue. One possible reason is that the training process is selected by two values: the canonical cost function in the model and the F1 score for feature selection, which intrinsically prevents overfitting. The low bias is probably because the Random Forest and AdaBoost methods are generally high-variance techniques and SVM can also exhibit high variance depending on choice of kernel.

Another observation is that both Random Forest and AdaBoost exhibit nondeterminism, meaning the F1 scores can vary across trials. Due to this random nature and the mild improvement of F1 scores during feature selection, there are substantial variations during the selection process and some features may be erroneously pruned just based on noise.

One final observations was that the learners prefer recall to precision, which is desired as this project is the first trial to identify possible diabetic patients, so high recall is better than high precision to ensure low false negative rate.

Discussion and Parameter Selection

To obtain the best model, we also performed parameter selection on logistic regression (Figure 4), AdaBoost (Figure 5) and SVM (Table 1).

For logistic regression, we fine-tuned the decision boundary by altering the cutoff probability. Lowering the cutoff probability increases recall and decreases the precision, which is preferred. The best recall score is about 0.86 with a cutoff probability of 0.32. If consider F1 score, the optimal is reached with a cutoff probability of 0.52 (Figure 4).

For the AdaBoost model using Decision Tree, there were three parameters to train: Max-Height - the maximal height of the Decision Tree, N-Estimator - the maximum number of estimators at which boosting is terminated, and Learning-Rate - the rate that each classifier's contribution shrinks. Since there is a trade-off between N-Estimator and Learning-Rate, we only optimized on Max-Height and N-Estimator. The selection result is shown in Figure 5a. It should be noted that boosting with SVM could also have been tried, as this method has been shown to address imbalanced datasets fairly well [Wang and Japkowicz (2008)].

We found that the selection process fit the training and validation sets much better than the test set (0.9 vs 0.4 F1 scores), and as the Max-Height increases, the F1 scores for training set and validation improved steadily while decreasing test set (especially recall), suggesting overfitting.

Table 1. SVM Results

Method	Kernel	Regularization Coefficient	Precision	Recall	F1 Score
SVM	RBF	0.1	0.1964	0.5431	0.2885
SVM	RBF	1.0	0.2212	0.5644	0.3183
SVM	RBF	10.0	0.1809	0.7110	0.2884
SVM	3rd Degree Polynomial	0.1	0.2127	0.6710	0.3230
SVM	3rd Degree Polynomial	1.0	0.3554	0.4510	0.3975
SVM	3rd Degree Polynomial	10.0	0.3554	0.4510	0.3975
SVM	Linear	0.1	0.2231	0.6543	0.3327
SVM	Linear	1.0	0.2231	0.6543	0.3327
SVM	Linear	10.0	0.2231	0.6543	0.3327
SVM	Sigmoid	0.1	0.1975	0.5414	0.2894
SVM	Sigmoid	1.0	0.2355	0.6968	0.3520
SVM	Sigmoid	10.0	0.2355	0.6968	0.3520

Table 2. Random Forest and AdaBoost Feature Selection Results

Method	Precision	Recall	F1 Score
RF Bottom-Up Under-Sampling	0.2894	0.7710	0.4208
RF Top-Down Under-Sampling	0.3333	0.4809	0.3938
RF Bottom-Up Over-Sampling	0.3345	0.7303	0.4588
RF Top-Down Over-Sampling	0.3307	0.7430	0.4577
Ada Bottom-Up Over-Sampling	0.3484	0.7277	0.4712
Ada Top-Down Over-Sampling	0.3515	0.6896	0.4656

This was suspected to be caused by the inconsistency between training set and test set as the training and validation sets were oversampled, while the test set is skewed. Thus, the skewed raw data was also used as training and validation sets to perform the same selection (Figure 5b).

The search on raw dataset, as expected, resulted in worse results, suggesting that balancing the data significantly improved the model, and we also noticed the same overfitting problem with high Max-Height. However, unlike the situation of the oversampled data, the search on raw data actually selected high precision, not recall, which is not desired. The possible reason is that the model favors labeling every patient as non-diabetic to reduce the training cost since the fraction of diabetic patients is low. Due to the observed complexity of parameter selection, possible future directions can include more sophisticated data balancing techniques, as this was a significant factor in the results.

We also found that the parameters with Max-Height of 1 and N-Estimator of 10 gave the best result on test data, reaching a precision of 0.3313 and a recall of 0.8244. Since these parameters favor a relatively low-variance learner, we felt that the selected features probably cannot accurately predict T2DM, and leaving relatively low correlation between features and T2DM occurrence. This is consistent with the fact that the data are full of invalid values and needs to be cleaned.

Among all the learners, logistic regression worked best, likely because the cleaned data set with abnormal records pruned was used. For the logistic regression model, most coefficients are quite small, even though they are significant. Lasso regularization was chosen to introduce sparsity in coefficients. This served as an auto feature-selection procedure. The best regularization parameter (the best lambda) was determined by cross-validation. The final model eliminated two features compared to the original logistic regression model. The lasso regularization has improved F1 score by 4% compared to the original model.

Conclusion and Future Work

In this project, four learning techniques were explored to predict T2DM. We made the following observations after significant analysis:

1. Balancing the data set can improve the prediction, and oversampling generally works better than undersampling.
2. The data set is highly diverse and contained significant amounts of invalid entries. Preprocessing is the key as logistic regression with the cleaned data set report the best performance.
3. The AdaBoost Model with decision tree works best with the un-cleaned data set. But minimal-height decision tree gave the best results, suggesting the very loose correlation between features and labels. Again, this proves data cleanup is essential.

In the future, we will focus on two aspects: a specifically designed cleanup method, and better feature selection through domain knowledge and other quantitative methods. We have shown that even adding a preliminary data cleanup step would significantly improve the generalized prediction accuracy, and there was low correlation between features and the true classification.

Acknowledgements

We would like to thank Kaggle for the project idea and also the dataset. Website:
https://www.kaggle.com/c/pf2012-diabetes. Most importantly, we want to thank the CS229 staff for the guidance and excellent instruction in Machine Learning to enable us to perform the analysis necessary for this project.

References

Causes of Diabetes (2014) *http://www.niddk.nih.gov/health-information/health-topics/Diabetes/causes-diabetes/Documents/Causes_of_Diabetes_508.pdf*, National Institute of Health

Diabetes Fact Sheet (2008) *http://www.who.int/nmh/publications/fact_sheet_diabetes_en.pdf*, World Health Organization

Mokdad A and Ford E (2003) *Prevalence of Obesity, Diabetes, and Obesity-Related Health Risk Factors*, The Journal of American Medical Association, VOl 289, No. 1

Practice Fusion Diabetes Classification (2012) *https://www.kaggle.com/c/pf2012-diabetes*, Kaggle

Wang B and Japkowicz N (2008) *Boosting Support Vector Machines for Imbalanced Data Sets*

Eschwege E, Richard JL, Thibult N, et al (1985)*Coronary heart disease mortality in relation with diabetes, blood glucose and plasma insulin levels. The Paris Prospective Study, ten years later*, Horm Metab Res Suppl, Vol 15

Statistical Fact Sheet (2014) *https://www.heart.org/idc/groups/heart-public/@wcm/@sop/@smd/documents/downloadable/ucm_462019.pdf*, American Heart Association

Drugs That Can Affect Blood Glucose Levels (2010) *http://www.diabetesincontrol.com/drugs-that-can-affect-blood-glucose-levels*, Diabetes In Control

M. Mihic, L. S. Mautner, J. Z. Feness, and K. Grant (1975) *Effect of trimethoprim-sulfamethoxazole on blood insulin and glucose concentrations of diabetics*, Can Med Assoc J., Vol 112

R. Chatterjee, H.C. Yeh, D. Edelman, and F. Brancati (2011) *Potassium and risk of Type 2 diabetes*, Expert Rev Endocrinol Metab., Vol 6