Preliminary Study of Longitudinal EMR data for Diabetes Forecasting

V. K. Potluru[1], P. Miller[2], J. Diaz-Montes[1], V. Hanagandi[2], J. Zola[3], M. Parashar[1],

[1] Rutgers Discovery Informatics Institute [2] Optimal Solutions Inc. [3] SUNY Buffalo

Motivation

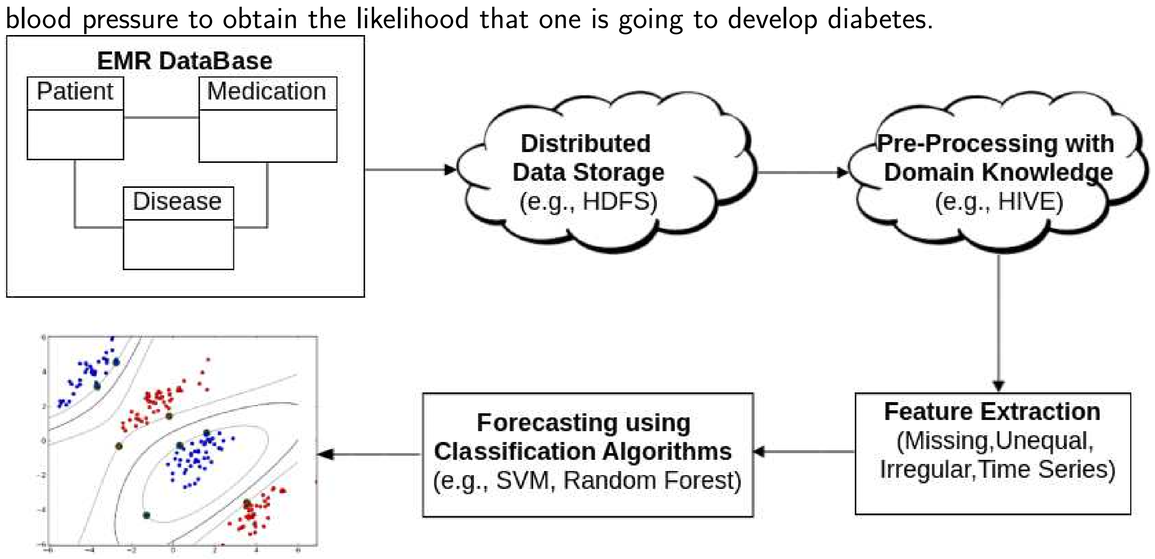
- Diabetes is a group of metabolic diseases in which there are high blood sugar levels over a prolonged period.
- Globally, in 2013, an estimated 344 million people have type 2 diabetes (T2D) [1].
- This is equal to 8.3% of the adults population, with equal rates in both women and men.
- Most symptoms of type 2 diabetes are not expressed aggressively and hence millions of patients remain undiagnosed for sustained periods of time.
- Health care providers and drug companies have an interest in identifying and treating these subjects before they develop full-blown diabetes.
- We would like to transform healthcare into heterogeneous and big data driven analytics for precise diagnosis and treatment

Problems

- Can we identify patients that are T2D positive from EMR data?
- How long prior to diagnosis can we accurately predict T2D?

Proposed Solution

We would like to use EMR data which routinely includes information such as age, BMI and blood pressure to obtain the likelihood that one is going to develop diabetes.



Preprocessing

We deployed a 2 TB of EMR dataset on Hive (SQL + Hadoop) with around 33 million subjects. To obtain positive and negative patients, we use the following criteria:

Positive patients

- Active problem instance of T2D ICD-9 codes
- Also require non-insulin anti-diabetic medication usage if Diabetes type unspecified
- Explicitly exclude subjects who are on Insulin

This resulted in 867,700 positive patients.

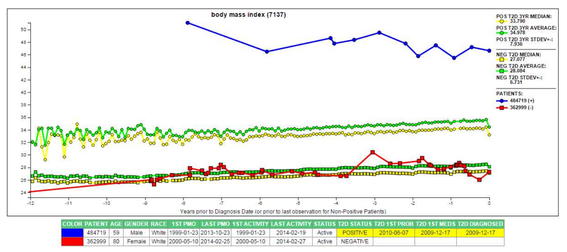
Negative patients

- Exclude subjects who received any anti-diabetic medications
- Exclude subjects who have any diabetes ICD-9 code (250.∗)
- Exclude subjects who have A1C values greater than 6 or glucose greater than 200
- Finally, exclude subjects who have one glucose value greater than 100

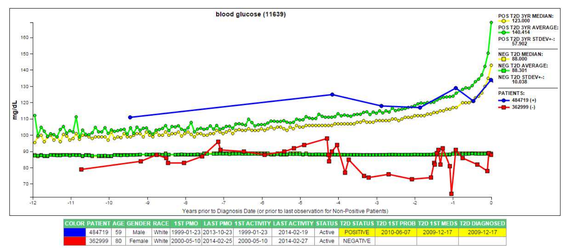
This gave us 3,056,666 negative subjects. We create two datasets for our experiments as follows.

- Diagnosis date for subjects may not be same as onset date.
- We take the minimum of onset date and medication start date in uncertain cases resulting in roughly half a million subjects.
- Smaller dataset of 10000 subjects consisting of 7 feature with rich time-series data.
- A larger dataset of a half a million subjects with 3 features to better match the real-world.

Visualization



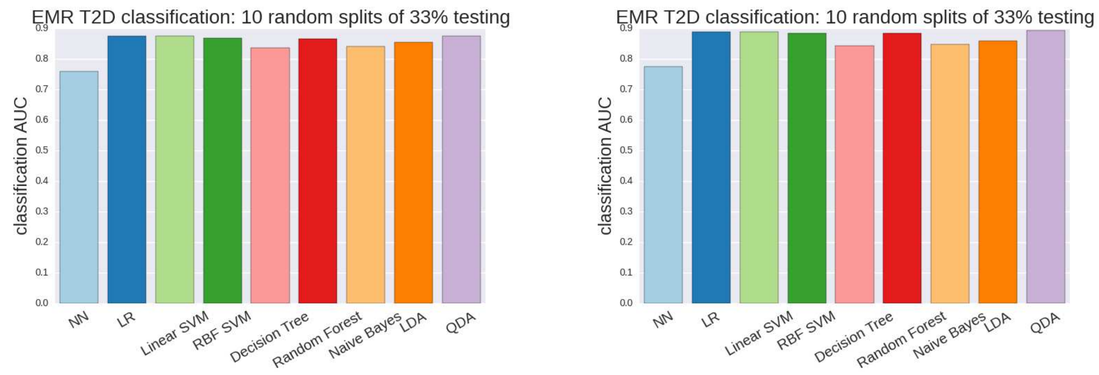
Trends for the BMI variable.



Trends for the glucose variable.

Experimental Results

- We then selected the 5000 patients which were most observation rich over our features.
- Various ML classifiers were used such as SVM, logistic regression, and random forests.
- AUC scores are shown using 10-fold cross validation.
- Prediction task consisted of data upto one and two years of diagnosis date.
- 7 features corresponding to BMI, systolic/diastolic blood pressure, A1C, creatinine, triglycerides and HDL.
- Derived features from the time-series — maximum, minimum, mean, standard deviation.



AUC scores using the various classifiers..

- AUC scores for logistic regression forecast at 0, 365 and 730 days.
- Half million subjects with training set being 10 percent of the total number of subjects.
- Logistic regression since it gave good performance on dataset of 10000 subjects.

Classifier	0 days	1 year	2 years
Logisitic Regression	0.82	0.72	0.74

Future work

- Experiment with other methods such as [2], [3] to see if we can get better performance.
- Predict T2D development in early stages by identifying patterns that lead to the disease development.

Acknowledgments

This research is supported by NSF-IIP-1346452.

References

[1] Simon Smyth and Andrew Heron.
Diabetes and obesity: the twin epidemics.
Nature medicine, 12(1):75–80, 2006.

[2] Jenna Wiens, Eric Horvitz, and John V Guttag.
Patient risk stratification for hospital-associated c. diff as a time-series classification task.
In *Advances in Neural Information Processing Systems*, pages 467–475, 2012.

[3] Jiayu Zhou, Fei Wang, Jianying Hu, and Jieping Ye.
From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records.
In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 135–144. ACM, 2014.