Analyzing DKA diagnosis after onset date using machine learning

Mari Shishikura[1,4], Justin Mower[2] PhD, Rona Sonabend[3] MD, Ila Singh[3] MD, Mark Rittenhouse[3], Devika Subramanian[2] PhD

[1]Faculty of Science, Division of Biological Science, Kyoto University, Japan
[2]Department of Computer Science, Rice University, U.S.A.
[3]Texas Children's Hospital, U.S.A.
[4]Nakatani RIES: Research and International Experiences for Student, Nakatani Foundation, Japan

Introduction

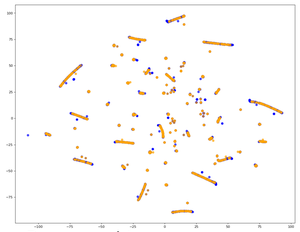
- Diabetic ketoacidosis (DKA) is a life-threating complication for children with Type 1 diabetes.
- More than 12,700 pediatric patients are diagnosed annually with DKA in the United States (with 2417 deaths in 2009), and are treated at a cost of over $90 million.
- Known risk factors for DKA include female gender, low socioeconomic status, ethnic minorities, and elevated A1C levels.
- In this research, we aim to use EHR data from Texas Children's hospital on Type 1 pediatric patients to build a comprehensive set of risk factors for accurately identifying patients at increased risk of DKA.
- By pro-actively identifying patients at high risk for DKA, we hope to intervene early to prevent DKA in Type 1 diabetic patients.

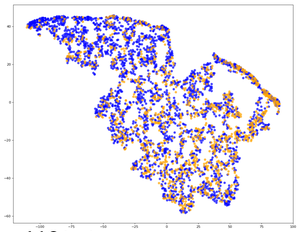
Methods

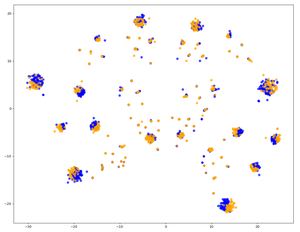
1. From the data provided by the Texas Children's Hospital, extract data about Type1 patients (total number of type1 patients = 4833).
    - Demographic data (Zip code, ethnicity etc.)
    - Lab data(A1C test, B-Hydroxybutyrate, C-peptide tests)
2. Separate the patients into 2 cohorts.
    - Cohort 1: Type 1 patients who have never had DKA, or no DKA after diagnosis. (3327)
    - Cohort 2: Type 1 patients who have had one or more DKAs after diagnosis. (1507)
3. Clean the data by omitting incomplete data and standardizing units for Lab data.
4. Create several sets of data that include different features for Cohort 1 and 2.
    - Demographics for cohort 1 and cohort 2
    - Lab data for cohort 1 and cohort 2
    - Demographics and Lab data for cohort 1 and 2
5. Use several classifiers and train each model with different data sets (in 5-fold CV).
    - Naïve Bayes
    - Logistic Regression (L1 and L2)
    - Random Forest Classifier
6. Evaluate the predictive accuracy of each classifier and also find the most important features for class separation.

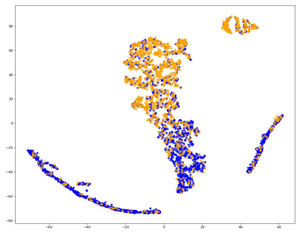
Results

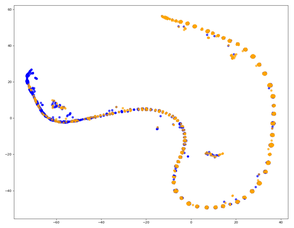
t- SNE plots (Blue: Cohort1, Orange: Cohort 2)

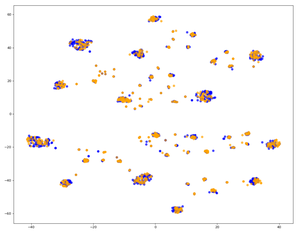

Demographics


A1C test


Demographics + A1C + B-Hydro + C-pep tests


B – Hydroxybutyrate test


C-Peptide test


Demographics + A1C +C-pep test

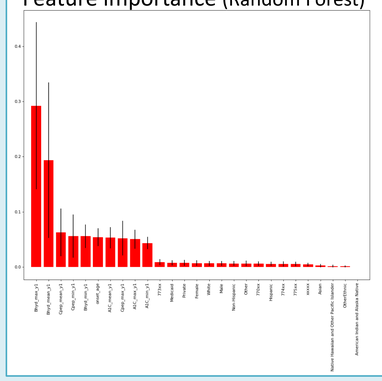
Information on B-Hydroxybutyrate and C-peptide appears to separate the two cohorts effectively. Here, we can tell that Demographics, A1C test, B-Hydroxybutyrate test, and C-peptide test values are optimal sets of information to be used for classification.

AUC values for different combinations of features and classifiers

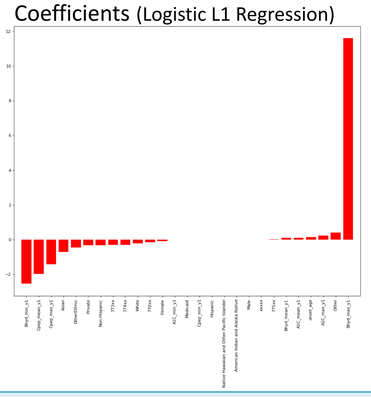
	Demographics	A1C tests	B-Hydroxybutyrate tests	C-peptide tests	Demographics + A1C + B-Hydro + C-pep tests	Demographics + A1C tests + C-peptide tests
Naïve Bayes	0.62	0.63	0.81	0.62	0.84	0.69
Logistic L2 Regression	0.62	0.64	0.82	0.72	**0.89**	0.72
Logistic L1 Regression	0.62	0.64	0.82	0.72	**0.89**	**0.74**
Random Forest Classifier (10 trees)	0.54	0.57	0.77	0.67	0.87	0.70
Random Forest Classifier (50 trees)	0.55	0.59	0.78	0.67	**0.88**	**0.74**

AUC values are indicators of how predictive the classifiers are. We can see that classifiers using information on all available features: Demographics, A1C test, B-Hydroxybutyrate test, C-peptide test, perform the best across all algorithms. In this case, the best models are L1 Logistic Regression and Random forests.

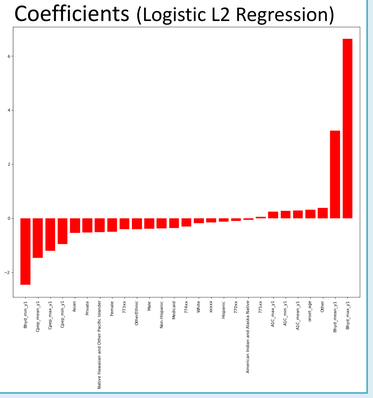
Feature importance (Random Forest)



Coefficients (Logistic L1 Regression)



Coefficients (Logistic L2 Regression)



Discussion

- B-Hydroxybutyrate tests are administered when doctors suspect a patient has DKA, so the values might not be useful in a predictive context.
- Patient's demographic information, onset age, A1C test values and C-peptide test values are the important discriminators.
- We showed that C-peptide test values are especially informative. C-peptide tests indirectly inform us of the levels of insulin produced in the body.
- Our work could assist physicians assess the risk of DKA in a patient (i.e., the probability of being in Cohort 1 or 2) and adjust therapeutic interventions appropriately, to reduce DKAs overall.
- The overall predictive AUC without B-Hydroxybutyrate reduces to 0.74 (from 0.88). We are working to improve this in our future analyses.
- These models have a limitation because they take in the overall lab data from 1st year after onset. It might not predict DKAs in the 1st year.

Future Work

- Train the classifiers with a larger number of patients. Some patients did not have complete data, which reduced the actual number used for training the models presented here.
- Add more features to the models. There is a lot of information about patients that we have not yet used –information on clinic encounters, pharmacy refills, visits to other units (co-morbidities), results from many more diabetes tests and panels.
- We have time series data on the lab values, so algorithms that take time index into account are an important next step in our analysis.

Acknowledgement

This research project was conducted as a part of the Nakatani Foundation's 2018 Nakatani RIES Fellowship for Japanese Students. For more information, visit http://nakatani-ries.rice.edu/.
Special thanks to the members of the Subramanian Group for their research mentorship and support and to Prof. Junichiro Kono, Sarah Phillips, and Aki Shimada. I am also thankful for people at Texas Children's Hospital for making this collaboration possible.

Reference

- Greko, F., Sather, R., *C-Peptide(Blood)*[Health Encyclopedia], University of Rochester Medical Center, Retrieved Sept 11, 2018, https://www.urmc.rochester.edu/encyclopedia/content.aspx?contenttypeid=167&contentid=c_peptide_blood
- Vladimir, S., Sherri, I. (2011) Role of beta-hydroxybutyric acid in diabetic ketoacidosis: A review, *Can Vet J*, 52(4):426-430
- Texas Children's Hospital, Texas Children's Take the Reins in Preventing DKA in High Risk Pediatrics Patients[Health Catalyst], Retrieved Sept 12, 2018, https://www.healthcatalyst.com/success_stories/dka-risk-prediction-texas-childrens-hospital
- Greko, F., Turley, R., Walton-Ziegler, O., *A1C*[Health Encyclopedia], University of Rochester Medical Center, Retrieved Sept 11, 2018, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3799221/
- Usher-Smith, J. A., Thompson, M. J., Sharp, S. J., Walter, F. M. (2011). Factors associated with the presence of diabetic ketoacidosis at diagnosis of diabetes in children and young adults: A systematic review. *The BMJ*. doi: 10.1136/bmj.d4092
- Schwartz, D. D., Axelrad, M. E., Anderson, B. J. (2014). A psychosocial risk index for poor glyfycemic control in children and adolescents with type 1 diabetes. *Pediatric Diabetes, 15*(3), 190-197. doi: 10.1111/pedi.1208