Delineating Pediatric Type 1 Diabetes Cohorts with Machine Learning

Mariko Mizogami,[1,2] Justin Mower,[3] Rona Sonabend,[4], Ila Singh,[4] Mark Rittenhouse,[4] Devika Subramanian[3]
[1]Department of Industrial and Management Systems Engineering, Waseda University, Tokyo, Japan
[2]2019 TOMODACHI STEM @ Rice University Program
[3]Department of Computer Science, Rice University, Houston, TX, U.S.A.
[4]Texas Children's Hospital, Houston, TX, U.S.A.

WASEDA University

RICE

What is DKA?

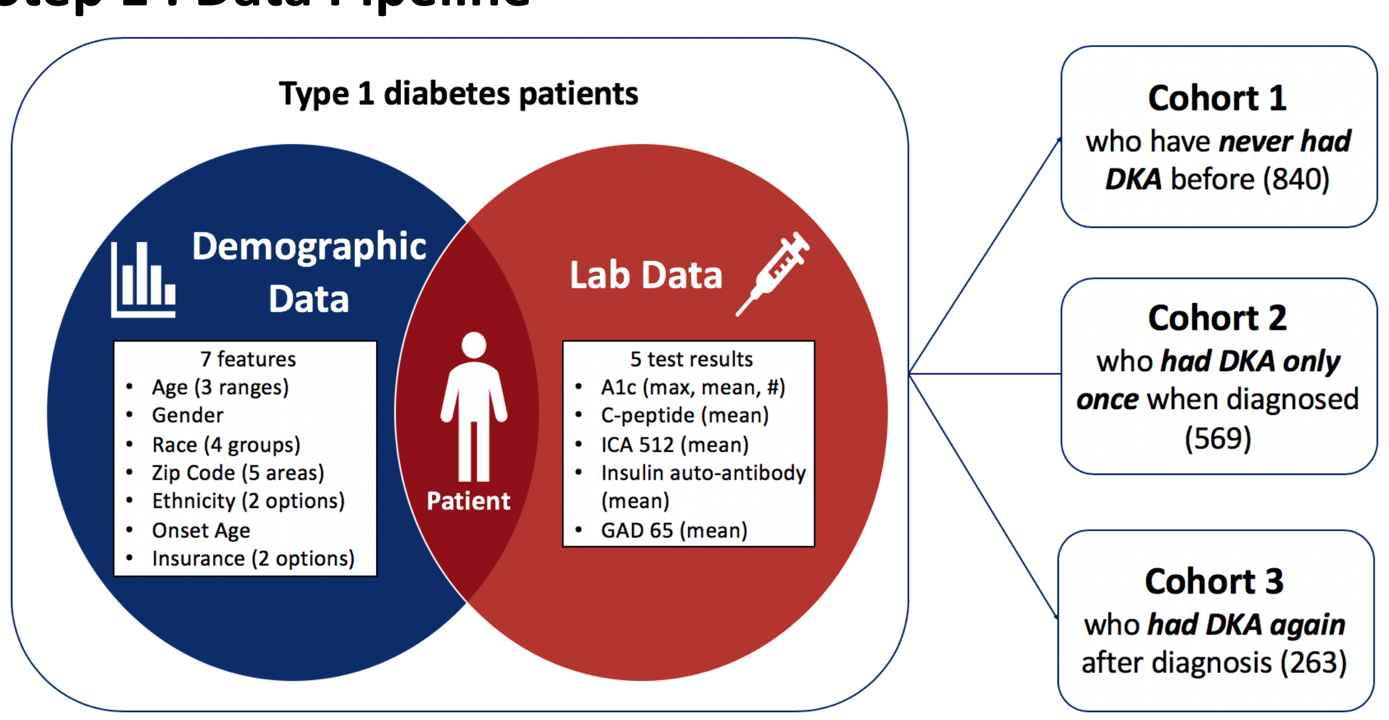
- Diabetic Ketoacidosis (DKA) is a preventable but life-threatening complication from Type 1 diabetes.
- Type 1 Diabetes is an early onset/juvenile diabetes.
  - More than 86,000 children are diagnosed with Type 1 diabetes every year.
- DKA is caused by a lack of insulin that causes high levels of blood acids called ketones.
  - Ketones poison the blood.
  - Can lead to coma and even death.

Data and Methods

Overview

- Collaborative research project with Texas Children's Hospital.
- Utilizes anonymized electronic medical record (EMR) data from pediatric patients.
- Goal is to build a model which can predict whether a type 1 diabetes patient is likely to have DKA in the future.

Step 1 : Data Pipeline

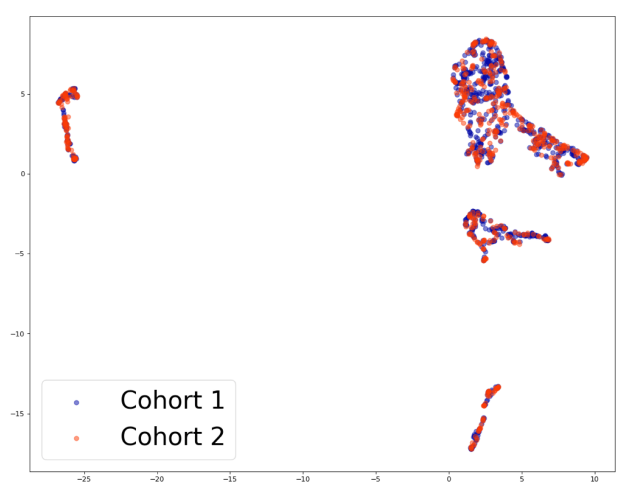


Step 2 : Machine Learning Analysis

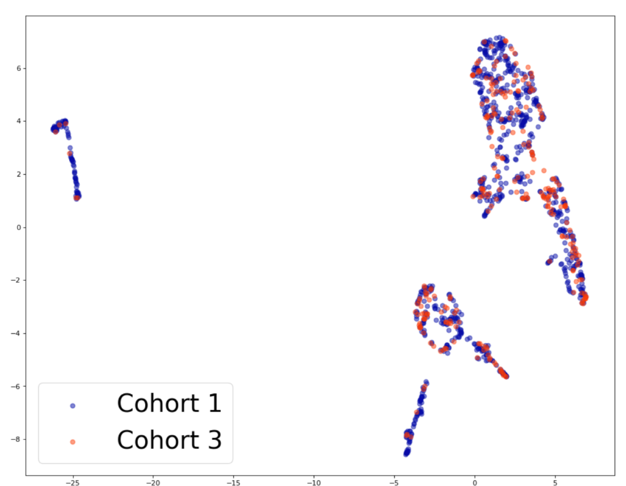
- Build 3 types of models for 3 combinations using lab and demographic data.
  - Logistic Regression, Random Forest, Gradient Boosting
  - 1 vs. 2, 1 vs. 3, 2 vs. 3.
- Plot uniform manifold approximation and projection (UMAP) for each combination, excluding categorical data.
- For each model, calculate area under the curve (AUC) and show important features.

Results

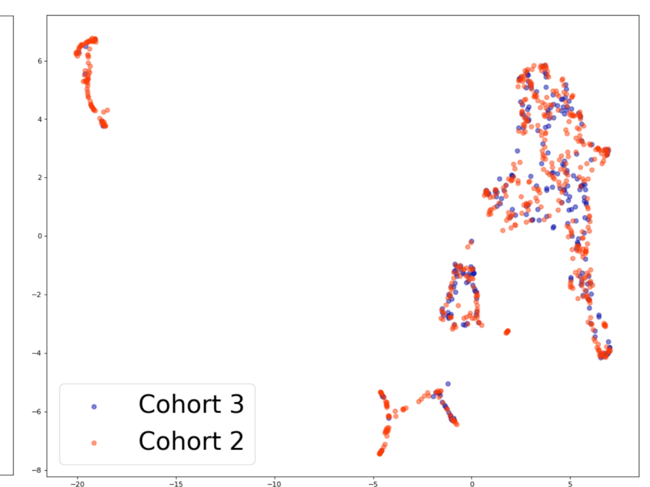
Pairwise Cohort UMAP Visualization



Cohort 1 vs Cohort 2



Cohort 1 vs Cohort 3



Cohort 2 vs Cohort 3

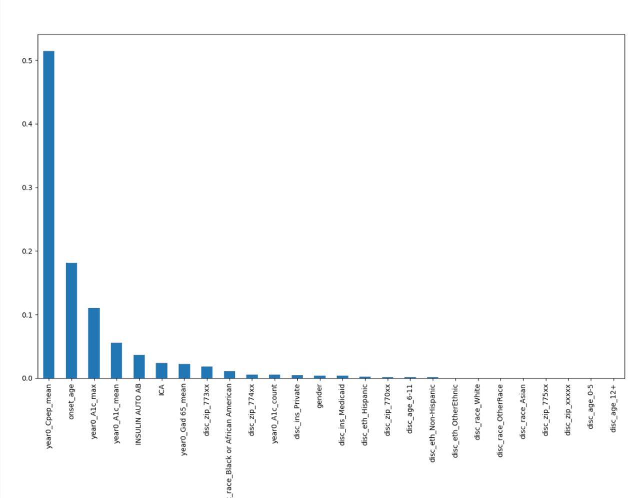
- All UMAP plots depict one large group with ≈3 auxiliary clusters.
- No matter which pair is being considered, the groups are not clearly and cleanly separable.

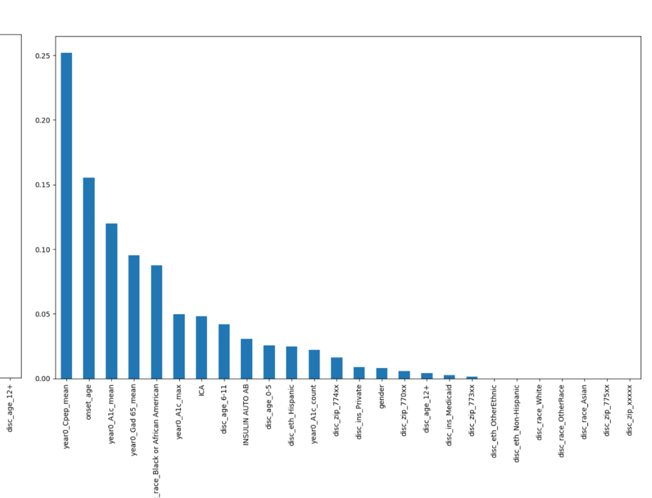
Pairwise Cohort Classification

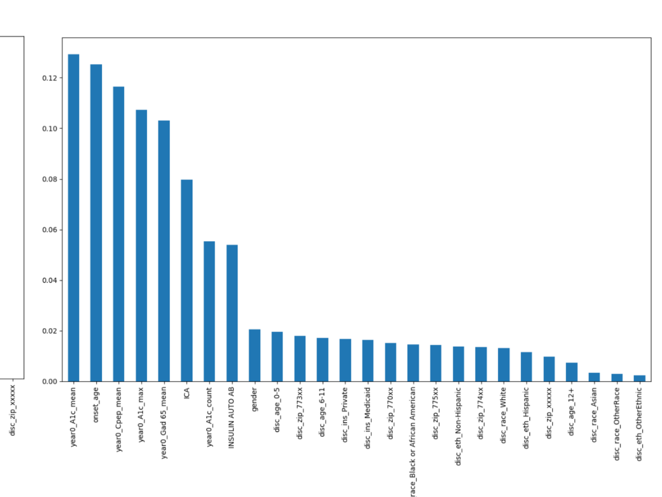
	Model Type	AUC	#1 Feature	#2 Feature	#3 Feature
Cohort 1 vs. Cohort 2	Gradient Boosting	0.80	C-pep mean (Ch1: 0.77 , Ch2: 0.39)	Onset Age (Ch1: 10.06, Ch2: 9.67)	A1c max (Ch1: 10.75, Ch2: 11.71)
Cohort 1 vs. Cohort 3	Gradient Boosting	0.78	C-pep mean (Ch1: 0.77, Ch3: 0.47)	Onset Age (Ch1: 10.06, Ch3: 9.59)	A1c mean (Ch1: 8.35, Ch3: 8.91)
Cohort 2 vs. Cohort 3	Gradient Boosting	0.66	GAD 65 mean (Ch2: 52.34, Ch3: 19.25)	A1c mean (Ch2: 8.80, Ch3: 8.91)	C-pep mean (Ch2: 0.39, Ch3: 0.47)

- C-pep mean, onset age, and A1c mean/max are generally the most predictive features for cohort classification.
- C-pep mean values show graded separation for each cohort consistent with clinical expectations.
- For the Cohort 2 vs. Cohort 3 model, GAD 65 mean was the most important.

Important Features for Each Combination


Cohort 1 vs. Cohort 2


Cohort 1 vs. Cohort 3


Cohort 2 vs. Cohort 3

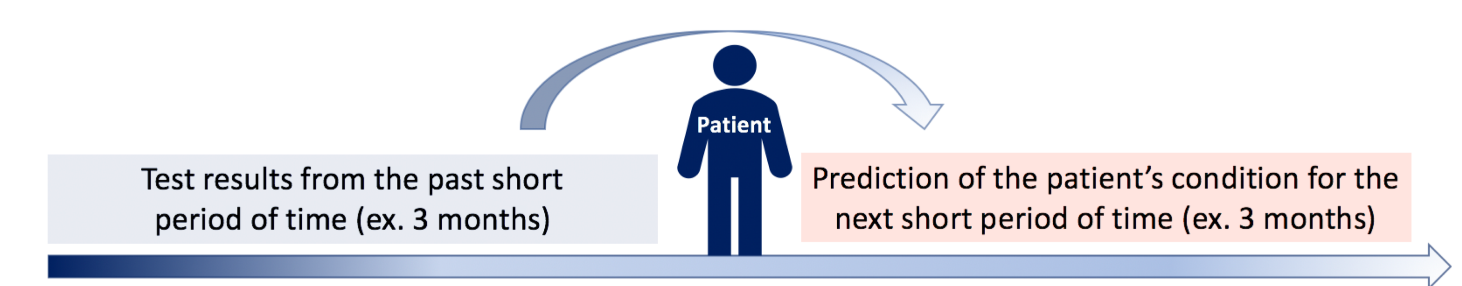
- Lab data is relatively more important than demographic data.
- Models separating cohorts 2 and 3 utilize more features than models separating the other cohorts.

Discussion and Conclusion

- UMAP plots suggest a generally consistent structure across cohorts.
- UMAP plots demonstrate the difficulty of classifying each cohort.
- Cohorts 2 and 3 are the least separable by these methods.
  - Both have DKAs in their first year after diagnosis, the only data used in this analysis.
- Each cohort has differing numbers of patients; balancing training sets may yield improvements.
- Additional data from BMI and blood pressure, as well as encounter data, wasn't included in this analysis.
- It is still unclear what features shape the subsets in each UMAP cluster.
- The features defining each cohort make clinical sense, and could help support physicians with diagnosis / prognosis

Future Research

1. Create ensemble models to improve performance.
2. Investigate UMAP clusters, revealing their possible clinical significance (i.e. are they clinically relevant subgroups of Type 1 diabetes).
3. Build a time-sensitive model, operating over the full range of time-indexed data we have available.



Acknowledgements

This research was conducted as part of the 2019 TOMODACHI STEM @ Rice University Program funded by the U.S.-Japan Council's TOMODACHI Initiative and with support by Dow Japan. I would like to thank the members of the Subramanian Group for their research mentorship. I would also like to thank TOMODACHI STEM program faculty, staff, and my fellow participants for their support and encouragement. For more information, visit http://tomodachistem.rice.edu/.

References

• 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 glyƒcemic control in children and adolescents with type 1 diabetes. Pediatric Diabetes, 15(3), 190-197. doi: 10.1111/pedi.1208.