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 Subramanian3

1Department of Industrial and Management Systems Engineering, Waseda University, Tokyo, Japan
22019 TOMODACHI STEM @ Rice University Program
3Department of Computer Science, Rice University, Houston, TX, U.S.A.
4Texas Children’s Hospital, Houston, TX, U.S.A.

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
  - 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

Results
- 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.
- Models separating cohorts 2 and 3 utilize more features than models with clinical expectations.
- 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

Discussion and Conclusion
- UMAP plots suggest a generally consistent structure across cohorts.
- UMAP plots demonstrate the difficulty of classifying each cohort.
  - Both have DKA's 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.
- 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
- N. J., and could help support physicians with diagnosis / prognosis