Predicting Future Type 1 Diabetes Onset Risk: A Machine Learning Approach

Hood D[1], Walia K[1], Kshirsagar O[1], Poddar S[1], Mankin B[1], Lee KCS[2], Adamek L[3], Rufino B[3], Josleyn J[4]

[1]Axtria, Berkley Heights, NJ, USA;
[2]Sanofi, Paris, France;
[3]Sanofi, Toronto Canada;
[4]Sanofi, Chicago, IL, USA

E-Poster

Copies of this poster obtained through Quick Response (QR) Code are for personal use only

POSTER HIGHLIGHT: Novel BERT Transformer Model Predicts Symptomless T1D Earlier, Potentially Reducing DKA Risk and Enabling Early Intervention

INTRODUCTION

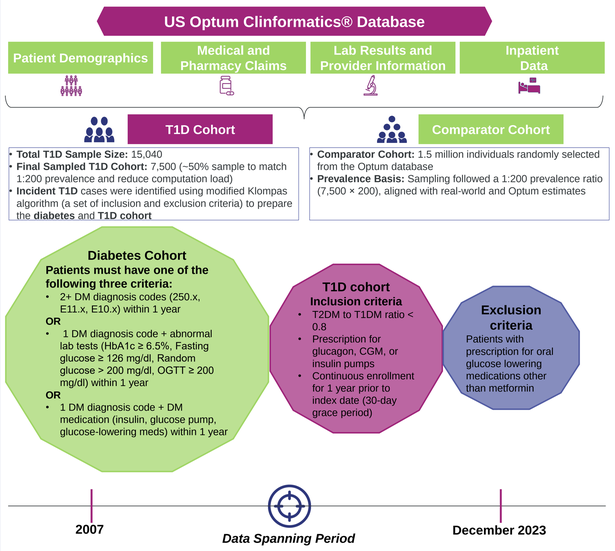
- Type 1 Diabetes (T1D) is an autoimmune condition that affected approximately 2 million people in the United States in 2021, including around 304,000 children and adolescents (aged <20 years) and 1.7 million adults (aged ≥20 years).[1]
- In the general population, the lifetime risk of developing T1D is about 1 in 200, but this risk increases to 1 in 20 among individuals with a family history, representing a 15-fold increase compared to 1 in 300 for those without.[2]
- Early identification of pre-symptomatic T1D reduces the risk of diabetic ketoacidosis (DKA) at diagnosis, with reported rates as low as 2.5%–5% in screened populations, enables timely intervention, and gives patients and families time to prepare for effective disease management, potentially improving long-term outcomes and reducing healthcare burden.[3-5]
- Traditional prediction methods rely on risk factors like family history and age, despite 80-90% of new cases having no family history.
- Machine learning (ML) uses real-world data and advanced algorithms to target more accurate predictions, enhancing the potential for early detection and intervention.[6,7]

OBJECTIVE

The objective of this research is to explore the potential of ML models in predicting the onset of T1D using US claims data and to identify the earliest timeframe at which T1D can be accurately predicted using ML techniques.

METHODS

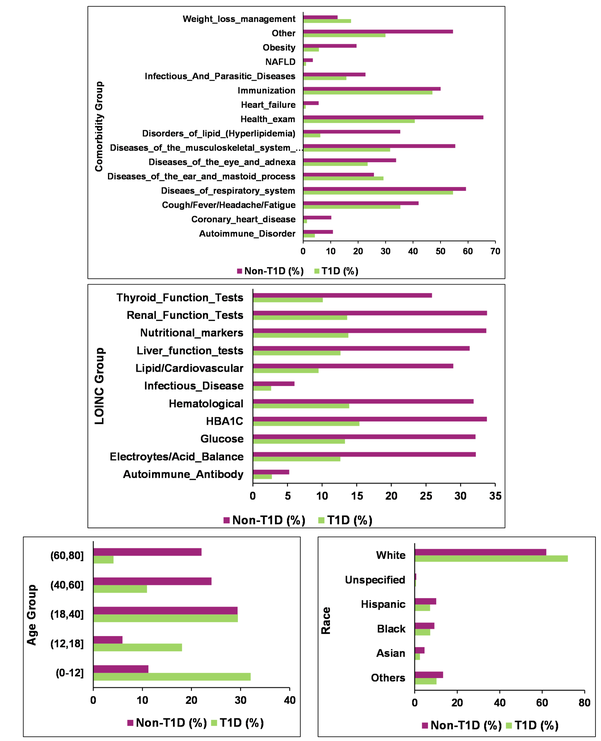
Figure 1: Study design and data sources



METHODS (cont'd)

- Figure 1 details the cohort creation methodology from the US Optum Clinformatics® Database[8], comparing T1D and comparator groups.
- Feature selection reduced 256 one-hot encoded features to 50 grouped variables for efficiency, with LASSO regression retaining all 50 variables.
- Figure 2 shows the input model features: including comorbidities, age group, race, and LOINC lab results for T1D and Non-T1D cohort.
- Various ML models (random forest, decision tree, logistic regression, XGBoost) were tested for predicting early diagnosis of pre-symptomatic T1D.
- Post hoc analysis using an optimized BERT model pre-trained completely on Optum dataset (1 year of medical event data), following the Med-BERT[9] architecture, and fine-tuned on three cohorts to enhance prediction accuracy for diabetes-related endpoints.
- This model was also applied to US claims data (T1D and non-T1D cohorts), utilizing key variables such as ICD, NDC, CPT codes, and demographics (age, race, gender, index date).

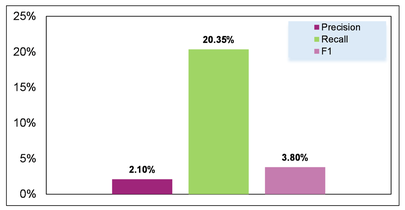
Figure 2: Key input features for T1D and non-T1D cohort post feature selection and engineering



RESULTS

- Extreme Gradient Boosting (XGB) with a class weight of 200 for T1D and a prediction threshold of 0.8 effectively addressed class imbalance, achieving precision: 2.1%, recall: 20%, F1 Score: 0.04, and Bayes Factor (BF): 4.67 (Figure 3).
- Applying a BF of 4.67 to a baseline rate of 1:200 changed it to 1:43, and for test data with a base rate of 0.9178 in 200, it became 1:47, equating to 448 true positives out of 21,362 predicted positives (Figure 3).
- Accuracy generally decreased as the time window moved from the index date, with an "elbow" in the performance curve identified as the optimal time window (12-24 months before the index date) for early detection, balancing early prediction and accuracy (Figure 4).
- The most important variables used by the model to make a T1D prediction are age (17.79%), COM_HEALTH_EXAM (12.40%), comorbidities including: hyperlipidemia (7.62%) and musculoskeletal disorders (6.02%).

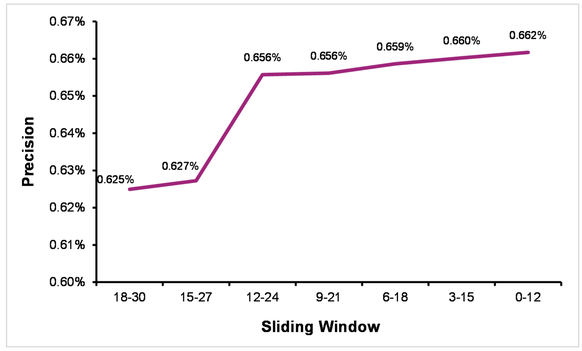
Figure 3: XGB model with W =200 and T =0.8 (12-24 months pre-index)



BF = 4.67	XGB
TP	448
FP	20,914
FN	1,753
TN	458,695

TP, true positive; FP, false positive; FN, false negative; TN, true negative

Figure 4: Sliding window results

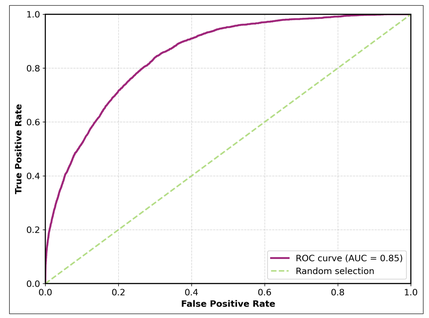


RESULTS (cont'd)

Table 1: Performance metrics for BERT model

Class/Metric	Precision	Recall	F1 Score
Non-T1D	99.78%	70.38%	82.54%
T1D	**2.59%**	**83.42%**	**5.03%**
Accuracy			70.51%
Macro Average	51.18%	76.90%	43.78%
Weighted Average	98.87%	70.51%	81.82%

Figure 5: ROC curve for BERT model



- The BERT model achieved an overall F1 score of 81.82% across the dataset, with excellent precision for the non-T1D cases (99.73%) and higher precision for T1D than XGB (2.59% vs 2.10%) (Table 1).
- The model achieved high recall for T1D (83.42%) (Table 1).
- The ROC curve shows an AUC of 0.85, indicating good overall discriminatory ability between T1D and non-T1D cohort (Figure 5).

CONCLUSION

- The study successfully demonstrated that significant predictive signals exist within the data, allowing for the development of effective and reliable models for early T1D detection.
- Future work will further explore OSCAR, a BERT derivative, incorporating new capabilities and broader data sources including lab tests, clinical procedures, imaging, genomics, wearables, and SDoH-related data.
- Exploring cost/benefit assumptions can refine model thresholds and weightings, ensuring accuracy and alignment with real-world clinical and economic contexts.

REFERENCES

1. Centers for Disease Control and Prevention (CDC). (2024, May 15). National Diabetes Statistics Report: Diabetes. Retrieved from https://www.cdc.gov/diabetes/php/data-research/?CDC_AAref_Val=https://www.cdc.gov/diabetes/data/statistics-report/index.html
2. TrialNet. (2025, April 15). Type 1 Diabetes Facts. Retrieved from https://www.trialnet.org/t1d-facts
3. Simmons, K. M., & Sims, E. K. (2023). Screening and prevention of type 1 diabetes: Where are we? The Journal of Clinical Endocrinology & Metabolism, 108(12), 3067–3079.
4. Schneider, J., Gemulla, G., Kiess, W., Berner, R., & Hommel, A. (2023). Presymptomatic type 1 diabetes and disease severity at onset. Diabetologia, 66(12), 2387–2388.
5. Hekkala, A. M., Ilonen, J., Toppari, J., Knip, M., & Veijola, R. (2018). Ketoacidosis at diagnosis of type 1 diabetes: Effect of prospective studies with newborn genetic screening and follow-up of risk children. Pediatric Diabetes, 19(2), 314–319.
6. Ghalwash, M., Anand, V., Ng, K., Dunne, J. L., Lou, O., Lundgren, M., et al. (2024). Data-driven phenotyping of presymptomatic type 1 diabetes using longitudinal autoantibody profiles. Diabetes Care, 47(8), 1424–1431.
7. Meng, W., Qin, J., Wang, T., & Zhao, R. (2025). Application of data science in management of type 1 diabetes. In Type 1 Diabetes – Causes, Treatments and Management. IntechOpen.
8. Optum Clinformatics® Database. (Accessed 2024). Available from: https://www.optum.com/business/solutions/life-sciences/real-world-data/claims-data.html
9. Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. NPJ Digital Medicine, 4(1), 86.

FUNDING

The study was sponsored by Sanofi.

ACKNOWLEDGMENT

Editorial assistance was provided by **Navneet Kumar, PhD**, **Shweta Shaw** and **Jatinder Kumar** of Axtria India Pvt Ltd. and was funded by Sanofi.

DISCLOSURES

Lee KCS, Adamek L, Rufino B, and **Josleyn J** are employees of Sanofi and may hold stocks /shares in Sanofi.

Hood D, Walia K, Poddar S, and **Mankin B** are employees of Axtria, Inc. and provided consulting services to Sanofi.

Kshirsagar O was an employee of Axtria at the time of the study