Principal Investigator

First Name: Eran
Last Name: Segal
Degree: Prof. (PhD)
Primary Affiliation: Weizmann Institute of Science
E-mail: Eran.Segal@weizmann.ac.il
State or Province: Israel
Country: Israel

General Information

Key Personnel (other than PI):
First Name: Hagai
Last name: Rossman
Degree: Ph.D.
Primary Affiliation: Weizmann Institute of Science
SCOPUS ID:
Requires Data Access? No

First Name: Gal
Last name: Sapir
Degree: M.D, Ph.D.
Primary Affiliation: Weizmann Institute of Science
SCOPUS ID:
Requires Data Access? Yes

First Name: Guy
Last name: Lutsker
Degree: MSc
Primary Affiliation: Weizmann Institute of Science
SCOPUS ID:
Requires Data Access? Yes

Are external grants or funds being used to support this research?: No external grants or funds are being used to support this research.
How did you learn about the YODA Project?: Internet Search

Conflict of Interest

https://yoda.yale.edu/wp-content/uploads/2024/02/Survey-Response.pdf
https://yoda.yale.edu/wp-content/uploads/2024/03/eran-coi.pdf
https://yoda.yale.edu/wp-content/uploads/2024/03/hagai-coi.pdf
https://yoda.yale.edu/wp-content/uploads/2024/02/COI-FORM-GL.pdf

Certification

Certification: All information is complete; I (PI) am responsible for the research; data will not be used to support litigious/commercial aims.
Data Use Agreement Training: As the Principal Investigator of this study, I certify that I have completed the YODA Project Data Use Agreement Training

1. NCT03267576 - 28431754DIA4026 - Canagliflozin Continuous Glucose Monitoring (CANA CGM) Trial: A Pilot Randomized, Double-Blind, Controlled, Crossover Study on the Effects of the SGLT-2 Inhibitor Canagliflozin (vs. the DPP-4 Inhibitor Sitagliptin) on Glucose Variability in Mexican Patients With Type 2 Diabetes Mellitus Inadequately Controlled on Metformin
2. NCT02139943 - 28431754DIA2004 - A Randomized Phase 2, Double-blind, Placebo-controlled, Treat-to-Target, Parallel-group, 3-arm, Multicenter Study to Assess the Efficacy and Safety of Canagliflozin as Add-on Therapy to Insulin in the Treatment of Subjects With Type 1 Diabetes Mellitus

What type of data are you looking for?: Individual Participant-Level Data, which includes Full CSR and all supporting documentation

Research Proposal

Project Title

Time-series insights into diabetes treatment - using a fine-tuned CGM foundation model to improve treatment outcomes

Narrative Summary:

Diabetes is a pressing health issue globally. We aim to investigate the potential of advanced learning approaches to improve the results of clinical trials by predicting treatment responses for individual patients. We will utilize proprietary data to enhance time-series based learning techniques in order to identify nuanced patterns in CGM data that correlate with treatment efficacy. Our research will focus on developing a methodological framework, paving the way for a deeper understanding of diabetes management. The anticipated outcome of our study is to establish a robust model capable of guiding clinical decisions, ultimately improving patient-specific diabetes treatment.

Scientific Abstract:

Background:
Diabetes, a rising health concern world-wide, demands improved treatment strategies. Advancements in Continuous Glucose Monitoring (CGM) and machine learning offer new avenues for personalized treatment approaches.
Objective:
To leverage advanced learning techniques, particularly time-series analysis, to predict individual treatment responses from CGM data in order to enhance diabetes treatment personalization and efficacy.
Study Design:
This research will employ proprietary CGM data from the Human Phenotype Project (HPP, previously called 10 K project (1). By integrating exploratory analysis tools like CGMap and IGLU and time-series based learning, we aim to refine predictive models for diabetes care.
Participants:
All participants who have more than 24 hours of CGM data available.
Primary and Secondary outcome measures:
Primary outcomes include the accuracy of treatment response predictions. Secondary outcomes are improved patient segmentation based on treatment efficacy and the identification of predictive markers for drug responsiveness.
Statistical analysis:
Our analysis will span descriptive, bivariate, and multivariable techniques, including machine learning models fine-tuned on the cohort data. We will evaluate model performance through cross-

validation and other relevant metrics, aiming to establish a robust framework for clinical decision support.

Brief Project Background and Statement of Project Significance:

Recent estimates suggest that around 6% of the world's population suffers from diabetes (~529 million people). This corresponds to approximately 38 million years of life lost due to roughly 1.7 million disease-related deaths (2). Additionally, it was recently estimated that ~9% of adults have impaired glucose tolerance, a number that is projected to increase significantly over the next two decades (3). Diabetes and related conditions are therefore major concerns for public health, worldwide.

Continuous glucose monitoring (CGM) devices measure the level of glucose in the interstitial fluid, which correlates well with blood glucose levels. The advantages of CGM usage in patients with diabetes was previously recognized (4), and recently stated to be an important component of future clinical trials concerning diabetic patients (5).

The 10 K cohort is a large-scale, prospective, longitudinal cohort of 40-70 years old Israeli individuals (1), containing, among other things, a 2-week CGM for each participant.

It was demonstrated that advanced statistical methods can improve our ability to describe diabetic patients (e.g. glucodensities (6)), and there are several available tools for the effective analysis of this data (e.g. (7),(8)). Additionally, it was shown that treating CGM data as time-series data can assist with treatment response prediction (9). Foundational models are able to generalize insights on new datasets that were not presented during training. Recent advancements have enabled this type of model for time-series data, and it is possible that such models could be used to enhance our understanding of diabetes through CGM data (10). We propose to apply these kinds of models on CGM data from clinical trials after it has been fine-tuned on sufficient data, in order to predict response to anti-diabetic treatments.

Specific Aims of the Project:

Predict Individual Treatment Responses: Utilize time-series analysis to predict how individuals with diabetes respond to specific treatments (DDP4 inhibitor, SGLT2 inhibitor).
Methodological Advancement: Develop and refine a methodological framework by integrating time-series based learning techniques with proprietary data from the 10 k cohort.
Patient Segmentation and Treatment Optimization: Identify patterns within CGM data that correlate with positive treatment outcomes, facilitating a nuanced segmentation of the patient population. This aims to optimize treatment strategies by aligning them more closely with individual patient profiles and needs.
Enhance Scientific and Medical Knowledge: By analyzing CGM data through the lens of advanced learning approaches, this project seeks to create new scientific knowledge that can be directly applied to enhance medical care for individuals with diabetes.

By focusing on individualized treatment response predictions and leveraging a rich dataset, we aim to advance the understanding and management of diabetes, with the potential to improve outcomes for millions worldwide.

Study Design:

Methodological research

What is the purpose of the analysis being proposed? Please select all that apply.

Develop or refine statistical methods

Research on clinical prediction or risk prediction

Research Methods

Data Source and Inclusion/Exclusion Criteria to be used to define the patient sample for your study:

Proprietary dataset from the 10k cohort study, including CGM data and comprehensive demographic and clinical information on participants.
Inclusion Criteria: All participants who have more than 24 hours of CGM data available.

Primary and Secondary Outcome Measure(s) and how they will be categorized/defined for your study:

Primary Outcome:
Patient Segmentation Efficiency: Evaluation of the model's ability to accurately segment patients based on their predicted response to treatment, assessed through clustering quality metrics.

Secondary Outcomes:

Predictive Accuracy: The precision and recall of treatment response predictions, determined through cross-validation and comparison with actual treatment outcomes.

Main Predictor/Independent Variable and how it will be categorized/defined for your study:

The primary predictor in our study will be the individual patient's CGM data and other demographic and clinical information. These variables will undergo quantitative analysis and categorization based on established clinical thresholds for glycemic control (e.g., HbA1c levels). They will form the dataset input for the model, influencing its output and serving as crucial indicators for predicting response to various diabetes treatments.

Other Variables of Interest that will be used in your analysis and how they will be categorized/defined for your study:

Other variables of interest include other demographic and clinical information. These variables will undergo quantitative analysis and categorization based on established clinical thresholds for glycemic control (e.g., HbA1c levels). They will form the dataset input for the model, influencing its output and serving as crucial indicators for predicting response to various diabetes treatments.

Statistical Analysis Plan:

Data preparation and Exploratory Analysis:
Data cleaning: standardize and apply transformation to skewed variables, address missing data.
Exploratory data analysis (EDA): conduct initial analysis to understand distributions, patterns and outliers. Visualize CGM time-series data, identify preliminary patterns of response to treatment. Additionally, perform covariate analysis to assess the influence of other demographic and clinical variables on treatment response.
Basic Analysis using CGMap and IGLU:
Utilize CGMap to calculate CGM metrics (mean glucose, glycemic variability etc).
Utilize IGLU for metrics including glucose management indicator (GMI) and other metrics.
Clustering for patient segmentation: Apply unsupervised learning techniques such as PCA, UMAP and HDBSCAN to segment patients into clusters based on similarities in CGM data patterns.
Utilize statistical tests to study inter-cluster differences in treatment response.
Treatment response prediction and Identification of Responders and Non-responders
Training a model for treatment response prediction by incorporating features extracted from CGMap and IGLU and a training set selected from the clinical trial data.
Using baseline measurements, predict the response to DDP4 inhibitor and SGLT2 inhibitor.

Using baseline measurements and one of the treatments, predict the response to the other treatment.
Using baseline measurements and the first X hours of a treatment, predict the final response to the treatment.
Advanced Time Series Characterization based on the TimeGPT foundation model:
Fine-tune the foundation model on a subset of the preprocessed CGM data from the 10 K cohort in order to build a CGM time-series model.
Feature extraction and time series analysis - extract relevant features such as time-series embeddings from CGM time-series data (possibly: complex temporal patterns, trends), then perform analysis on extracted features. For example: change in features due to treatment.
Train a model for treatment response prediction by incorporating the fine-tuned CGM time-series model and a training set selected from the clinical trial data, as described in (4a).
Model Evaluation and Validation
Compare the relation between metrics in feature space (3a, 5b) and treatment response. For example, distances between samples within a group who responded well to the treatment compared to the distance between groups of samples (responders and non-responders).
Utilize N-fold cross-validation to train a treatment response predictor (4a, 5c) on the clinical trial data. Evaluate based on ROCAUC, precision and recall.

Software Used:

Python

Project Timeline:

Months 1-2: Define goals and preprocess CGM data, perform EDA on trial data (basic CGM analysis - CGMap, IGLU).
Months 3-6: Fine-tune TimeGPT with CGM data.
Months 7-8: Extract features and enhance predictive models using TimeGPT.
Months 9-10: Perform advanced analysis and refine models.
Months 11-12: Finalize report and prepare manuscript for publication.

Key Milestones:

End of Month 2: Preprocessed dataset ready.
End of Month 6: TimeGPT fine-tuned.
End of Month 8: Predictive models enhanced.
End of Month 10: Model refinement and advanced insights complete.
Analysis Completion Date: End of Month 10.
Manuscript Drafted: Month 11.

First Submission for Publication: By Month 12.

Results Reported Back to YODA Project: By Month 12.

Dissemination Plan:

We aim to publish our findings in top-tier journals such as Nature Medicine or NEJM, targeting a broad audience across diabetes care and medical research. Plans include presenting at international conferences like ADA and EASD. By choosing open-access options, we ensure wide accessibility, benefiting healthcare professionals and fostering advancements in personalized diabetes management.

Bibliography:

1. Shilo S, Bar N, Keshet A, Talmor-Barkan Y, Rossman H, Godneva A, Aviv Y, Edlitz Y, Reicher L, Kolobkov D, Wolf BC, Lotan-Pompan M, Levi K, Cohen O, Saranga H, Weinberger A, Segal E. 10 K: a large-scale prospective longitudinal study in Israel. Eur J Epidemiol. 2021

Nov;36(11):1187-1194. doi: 10.1007/s10654-021-00753-5. Epub 2021 May 15. PMID: 33993378.

2. Watkins DA, Ali MK. Measuring the global burden of diabetes: implications for health policy, practice, and research. Lancet. 2023 Jul 15;402(10397):163-165. doi: 10.1016/S0140-6736(23)01287-4. Epub 2023 Jun 22. PMID: 37356449.
3. Mary R. Rooney, Michael Fang, Katherine Ogurtsova, Bige Ozkan, Justin B. Echouffo-Tcheugui, Edward J. Boyko, Dianna J. Magliano, Elizabeth Selvin; Global Prevalence of Prediabetes. Diabetes Care 1 July 2023; 46 (7): 1388–1394. https://doi.org/10.2337/dc22-2376
4. Danne T, Nimri R, Battelino T, Bergenstal RM, Close KL, DeVries JH, Garg S, Heinemann L, Hirsch I, Amiel SA, Beck R, Bosi E, Buckingham B, Cobelli C, Dassau E, Doyle FJ 3rd, Heller S, Hovorka R, Jia W, Jones T, Kordonouri O, Kovatchev B, Kowalski A, Laffel L, Maahs D, Murphy HR, Nørgaard K, Parkin CG, Renard E, Saboo B, Scharf M, Tamborlane WV, Weinzimer SA, Phillip M. International Consensus on Use of Continuous Glucose Monitoring. Diabetes Care. 2017 Dec;40(12):1631-1640. doi: 10.2337/dc17-1600. PMID: 29162583; PMCID: PMC6467165.
5. Battelino T, Alexander CM, Amiel SA, Arreaza-Rubin G, Beck RW, Bergenstal RM, Buckingham BA, Carroll J, Ceriello A, Chow E, Choudhary P, Close K, Danne T, Dutta S, Gabbay R, Garg S, Heverly J, Hirsch IB, Kader T, Kenney J, Kovatchev B, Laffel L, Maahs D, Mathieu C, Mauricio D, Nimri R, Nishimura R, Scharf M, Del Prato S, Renard E, Rosenstock J, Saboo B, Ueki K, Umpierrez GE, Weinzimer SA, Phillip M. Continuous glucose monitoring and metrics for clinical trials: an international consensus statement. Lancet Diabetes Endocrinol. 2023 Jan;11(1):42-57. doi: 10.1016/S2213-8587(22)00319-9. Epub 2022 Dec 6. Erratum in: Lancet Diabetes Endocrinol. 2024 Feb;12(2):e12. PMID: 36493795.
6. Matabuena M, Petersen A, Vidal JC, Gude F. Glucodensities: A new representation of glucose profiles using distributional data analysis. Statistical Methods in Medical Research. 2021;30(6):1445-1464. doi:10.1177/0962280221998064
7. Shao J, Liu Z, Li S, Wu B, Nie Z, Li Y, Zhou K. Continuous Glucose Monitoring Time Series Data Analysis: A Time Series Analysis Package for Continuous Glucose Monitoring Data. J Comput Biol. 2023 Jan;30(1):112-116. doi: 10.1089/cmb.2022.0100. Epub 2022 Aug 8. PMID: 35939283.
8. Broll S, Urbanek J, Buchanan D, Chun E, Muschelli J, et al. (2021) Interpreting blood GLUcose data with R package iglu. PLOS ONE 16(4): e0248560. https://doi.org/10.1371/journal.pone.0248560
9. Li L, Sun J, Ruan L, Song Q. Time-Series Analysis of Continuous Glucose Monitoring Data to Predict Treatment Efficacy in Patients with T2DM. J Clin Endocrinol Metab. 2021 Jul 13;106(8):2187-2197. doi: 10.1210/clinem/dgab356. PMID: 34010405.
10. Garza, Azul, and Max Mergenthaler-Canseco. “TimeGPT-1.” arXiv preprint arXiv:2310.03589 (2023).