140.777.01 – Statistical Programming Paradigms and Workflows
Final Write-Up: The Mouseketeer Dossier

All for One, One for Glucose

A Mouseketeer Approach to Glycemic Prediction

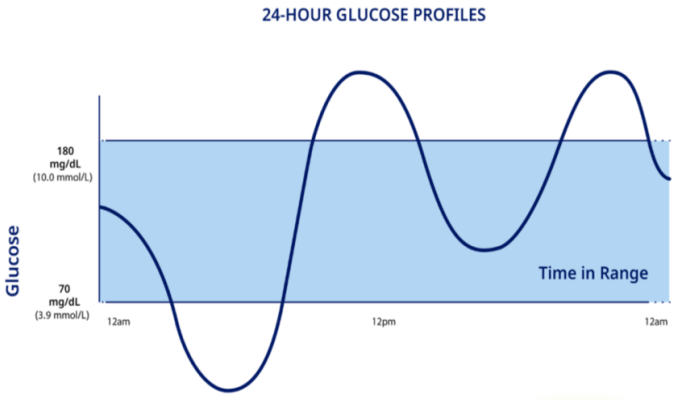
GitHub Link: https://github.com/jhu-statprogramming-fall-2025/project04-the-3-mouseketeers

The 3 Mouseketeers:	Connor Moore	cmoor150@jh.edu	
	Eimee Co	eco1@jh.edu	Esprit-Cat
	Jin Sung Sur	jsur3@jh.edu	jinsungsur

I. Introduction

Improvements in technology such as wearable health devices have allowed for more accurate and continuous monitoring of a patient's health status. One example is the use of continuous glucose monitors (CGMs) which patients with diabetes wear to track their blood glucose levels to assess blood sugar (glycemic) control. Some CGMs include a pump that automatically delivers insulin based on programmed settings and patient readings -all of which are recorded and transmitted electronically. Clinicians now have access to more data and opportunities than ever before, and it is imperative to leverage this resource to improve clinical decision making and improve health outcomes.

Time-In-Range (TIR), one measure of glycemic control, is defined as percentage time when glucose is between range **(70-180mg/dL).**[1] This is clinically relevant because it is a strong predictor of complications such as retinopathy and nephropathy. After considering both the benefits of tight control and practical feasibility, current guidelines recommend a **TIR ≥ 70% as the clinical target** for adequate control.



Given its impact on health, being able to easily flag or predict patients who are or are likely to be uncontrolled can help clinicians intervene through education, insulin adjustments or technological support, ultimately improving patient's health.

II. Objectives

This exploratory project aims to determine the feasibility of developing a classification or a predictive model for adequate glycemic control (TIR ≥ 70%) based on existing patient data: such as demographic information and insulin pump-derived metrics.

Using statistical workflows and paradigms learned in class, these are the objectives of our project:

1. **Data Frame Construction**: Identify, collect, and construct a clean and appropriate data set through functional programming and data wrangling techniques.

2. **Machine Learning Model:** Apply machine learning methods to build and evaluate classification and predictive models, using lasso for variable selection and generalized

linear models (GLMs) as a baseline for future models.

3. **Dashboard Deployment:** Deploy an interactive website which summarizes findings.

III. Data Frame Construction

a. Data Source and Access

The chosen dataset is from Brown's study on the use of different CGMs and insulin pumps among 168 children and adults with Type 1 Diabetes.[2] The study participants' ages ranges from 14 to 71 years and their long-term glucose control (measured by A1c) ranges from 5.4 to 10.6%.[2] This dataset is publicly available and can be accessed online by request: https://public.jaeb.org/dataset/573.[3]

The dataset includes 50 different data frames, with a cumulative 17 million rows and several hundred million individual data points**.** These are the specific data frames and examples of some variables we chose to include for this project:

- DiabScreening
- DiabSocioEcon
- DiabLocalHbA1c
- MedicalCondition
- Medications
- Pump_CGM Value
- Pump_BasalRateChange
- Pump_Bolus Delivered

b. Data Challenges

The dataset contained a lot of information, and it required time to understand and be familiar with what the different datasets and different variables represented. Relevant information was spread in different .txt files and there were many missing values.

c. Data wrangling

Data wrangling was not limited to constructing a data frame including the outcome of interest and all relevant possible predictors. We also created summary variables to make the data more useable for analysis and understandable for our users, such as:

- BMI (or Body Mass Index from height and weight)
- TDD (or Total Daily Dose)
- Basal and Bolus Statistics
- Percentage in Range (from discrete CGM readings and their time stamps)

d. Data Description

Here are some example descriptive statistics of the sample population:

- Pump cohort N: 125
- Mean TIR (70–180): 71.5%
- Median TIR (70–180): 71.4%
- Uncontrolled (% TIR < 70): 37.6%
- Mean HbA1c: 7.66
- Mean TDD (units/day): 54

These are the variables we ultimately chose to include in the analysis:

1. **Outcome: Uncontrolled**
   - a Binary variable derived from pct_70_180 which is percentage time when glucose reading was between 70 to 180 mg/dL).
   - 0 for values ≥ 70% and 1 otherwise.

2. **Predictors:**

Patient	Pump-Derived Metrics
Demographics • AgeAtEnrollment • Gender • Ethnicity • Race **Socioeconomic** • EducationLevel • AnnualIncome • InsuranceType **Diabetes Diagnosis** • DiagAge • DiagAgeApprox **Past Medical History/Medications** • NumMedicalConditions • NumMedications **Physical Examination Findings** • Weight_kg • Height_cm • BMI • BldPrSys • BldPrDia **Laboratory Results** • HbA1c_Screening	**Total Daily Dose (TDD) Metrics** • tdd_mean • tdd_median • tdd_sd • tdd_cv • TDD_per_kg • TDD_per_BMI **Basal insulin metrics** • basal_mean • basal_median • basal_sd • basal_pct **Bolus insulin metrics** • bolus_mean • bolus_median • bolus_sd • bolus_pct • mean_boluses_per_day **Data collection duration** • n_days

IV. Machine Learning Models[4]

```
packages <- c("caret", "glmnet", "pROC", "ggplot2")
lapply(packages, library, character.only = TRUE)
```

We decided to create two models for different functions: one including all predictors (Full model) and one excluding pump-derived metrics (No Insulin model).

The Full model will be useful for flagging existing patients for whom we already have insulin pump readings. The No Insulin model can be used for new patients with no existing pump data. It also accounts for potential leakage. For example, insulin (especially bolus doses) can be triggered by uncontrolled glucose (our outcome of interest).

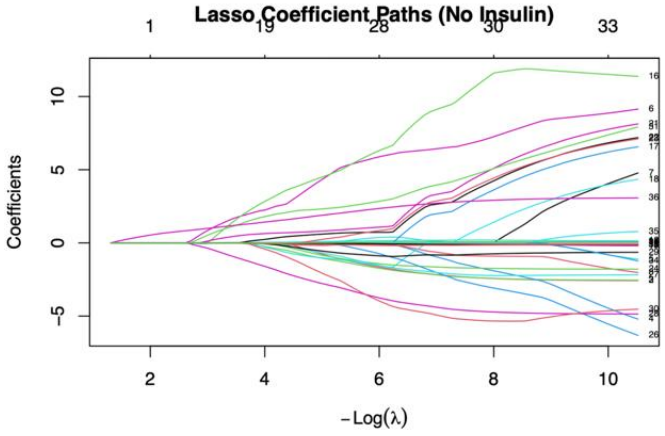
a. Data Processing

Minimal processing was needed to prepare the data for GLM. This included dropping the rows that did not have an outcome variable, imputing missing values using median, and transforming all characters into factors. These were then split into a training set and a test set in an 80:20 ratio, using the partition function to ensure comparable ratios of Controlled/Uncontrolled outcomes.

b. Ridge/Lasso for Variable Selection

Ridge and Lasso are methods for variable selection that It helps narrow down a wide range of variables to those that have the most impact on the chosen outcome by balancing predictive value with the tuning parameter that shrinks variables to either towards zero (Ridge) or to zero (Lasso). Of note, lasso is better for sparsity, though ridge is more stable. While we did both during the analysis, we decided to go with lasso for this model, as this is a preliminary project.

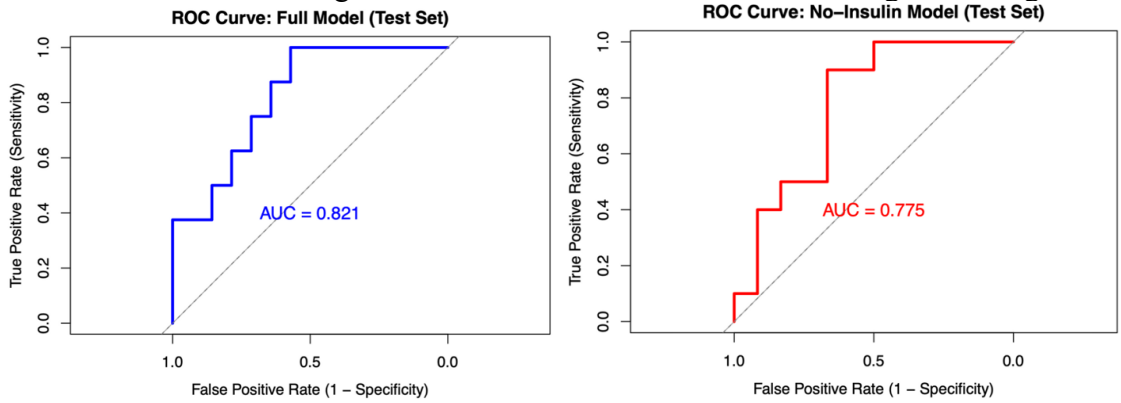
Other options for variable selection are subset selection and principal component analysis (PCA). Lasso is faster to run and accounts for collinearity when selecting variables, unlike subset selection. It is also easily understandable and interpretable, unlike PCA.



c. Model Fitting and Evaluation

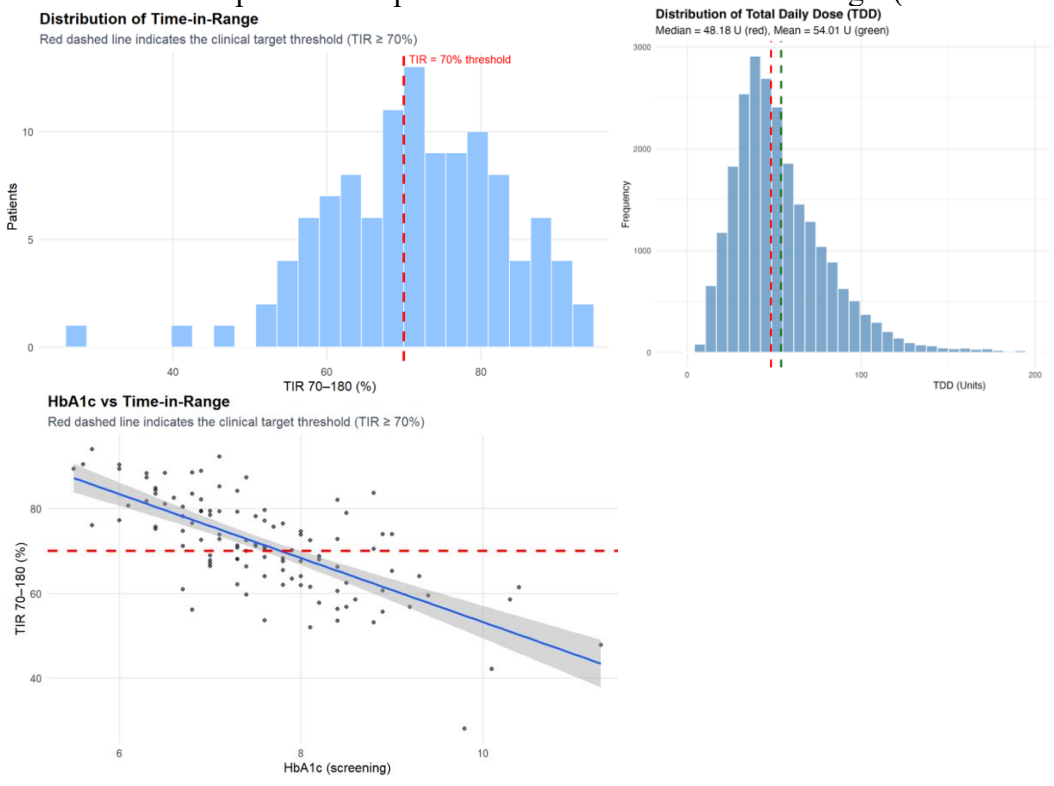
The variables selected by Lasso in both Full and No insulin were fitted into Generalized Linear Models (GLMs) using the training set.

Performance was assessed using confusion matrixes and AUC curves using the testing set.



V. Website Deployment

To communicate our findings, we launched a quarto dashboard, which includes graphs that summarize what we believe are important statistics, such as TIR, TDD, and HbA1c. It also has some interactive options to help users better understand our findings. (See Submitted Link)



VI. Going Forward

a. Insights

Clinical:

- IR < 70% is common even in a trial cohort
- HbA1c and insulin-delivery patterns are strongly associated with uncontrolled status.
- Even without insulin features, age and HbA1c still carry a meaningful signal.

Modelling:

- Lasso can select collinear variables and give a more manageable set of predictors
- Data leakage is a concern that needs to be addressed in predictive models
- ROC/AUC clarified performance beyond accuracy alone.

Workflow:

- Gluing together many R scripts is fragile; having a clear pipeline diagram and consistent file paths matters more than adding another fancy model.
- We also saw that helper functions and a single combined patient table made downstream modeling much easier to iterate on.

    b. Steps forward

- Fit to other models, such as decision trees, random forest, kernel
- Apply model/pipeline with other or larger datasets

140.777.01 – Statistical Programming Paradigms and Workflows
Final Write-Up: The Mouseketeer Dossier

References for R Packages used:

- Kuhn M. *caret: Classification and Regression Training.* R package version 6.0-93. 2023. https://CRAN.R-project.org/package=caret
- Friedman J., Hastie T., Tibshirani R. *glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models.* R package version 4.1-7. 2023. https://CRAN.R-project.org/package=glmnet
- Robin X., Turck N., Hainard A., Tiberti N., Lisacek F., Sanchez J.-C., Müller M. *pROC: Display and Analyze ROC Curves.* R package version 1.18.0. 2023. https://CRAN.R-project.org/package=pROC
- Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* R package version 3.4.3. 2023. https://ggplot2.tidyverse.org
- Broll S., Urbanek J., Buchanan D., Chun E., Muschelli J., Punjabi N., Gaynanova I. Blood glucose data with R package iglu. *PLOS ONE.* 2021;16(4):e0248560. doi:10.1371/journal.pone.0248560
- Chun E., Broll S., Buchanan D., Muschelli J., Fernandes N., Seo J., Shih J., Urbanek J., Schwenck J., Gaynanova I. iglu: Interpreting glucose data from continuous glucose monitors. R package version 3.5.0. 2023.
- Friedman J., Hastie T., Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software.* 2010;33(1):1–22. doi:10.18637/jss.v033.i01
- Tay J.K., Narasimhan B., Hastie T. Elastic net regularization paths for all generalized linear models. *Journal of Statistical Software.* 2023;106(1):1–31. doi:10.18637/jss.v106.i01
- Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, Markus Müller. PROC: An open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics.*2011;12:77. doi:10.1186/147121051277

References for Write-Up:

1. Battelino T., Danne T., Bergenstal R.M., et al. Clinical targets for continuous glucose monitoring data interpretation: Recommendations from the International Consensus on Time in Range. *Diabetes Care.* 2019;42(8):1593–1603. doi:10.2337/dc19-0028
2. Brown (2019), *The International Diabetes Closed Loop (IDCL) Trial: Clinical Acceptance of the Artificial Pancreas - A Pivotal Study of t:slim X2 with Control-IQ Technology (DCLP3), (Version 4)*, Retrieved from https://github.com/irinastatslab/awesome-cgm/wiki/Brown-(2019)
3. Brown S.A., Kovatchev B.P., Raghinaru D., et al. Six-month randomized, multicenter trial of closed-loop control in type 1 diabetes. *New England Journal of Medicine.* 2019;381(18):1707–1717. doi:10.1056/NEJMOA1907863
4. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning: With Applications in R* (2nd ed.). Springer; 2021. doi:10.1007/978-1-0716-1418-1

Other References:

- Anderson S.M., Dassau E., Raghinaru D., et al. The International Diabetes Closed-Loop Study: Testing artificial pancreas component interoperability. *Diabetes Technology & Therapeutics.* 2019;21(2):73–80. doi:10.1089/dia.2018.0308

- Beck R.W., Bergenstal R.M., Riddlesworth T.D., Kollman C. The association of biochemical hypoglycemia with the subsequent risk of a severe hypoglycemic event: Analysis of the DCCT data set. *Diabetes Technology & Therapeutics.*2019;21(1):1–5. doi:10.1089/dia.2018.0362

- Contreras I., Vehi J. Artificial intelligence for diabetes management and decision support: Literature review. *Journal of Medical Internet Research.* 2018;20(5):e10775. Published 2018 May 30. doi:10.2196/10775

- Kiran M., Xie Y., Anjum N., Ball G., Pierscionek B., Russell D. Machine learning and artificial intelligence in type 2 diabetes prediction: A comprehensive 33-year bibliometric and literature analysis. *Frontiers in Digital Health.*2025;7:1557467. Published 2025 Mar 27. doi:10.3389/fdgth.2025.1557467