Predicting Insulin Pump Therapy Settings

Riccardo L. Ferraro  
*Southern Methodist University & Tandem Diabetes Care, Inc,* riccoferraro@gmail.com

David Grijalva  
*Southern Methodist University,* grijalva789@gmail.com

Alex Trahan  
*Tandem Diabetes Care, Inc,* atrahan@tandemdiabetes.com

Follow this and additional works at: https://scholar.smu.edu/datasciencereview

Part of the Applied Statistics Commons, Biomedical Devices and Instrumentation Commons, Data Science Commons, Endocrine System Diseases Commons, Multivariate Analysis Commons, Other Statistics and Probability Commons, Signal Processing Commons, and the Statistical Models Commons

Recommended Citation

Ferraro, Riccardo L.; Grijalva, David; and Trahan, Alex (2022) "Predicting Insulin Pump Therapy Settings," *SMU Data Science Review.* Vol. 6: No. 2, Article 15.  
Available at: https://scholar.smu.edu/datasciencereview/vol6/iss2/15

This Article is brought to you for free and open access by SMU Scholar. It has been accepted for inclusion in SMU Data Science Review by an authorized administrator of SMU Scholar. For more information, please visit http://digitalrepository.smu.edu.
Predicting Insulin Pump Therapy Settings

Ricco Ferraro\textsuperscript{1,2}, David Grijalva\textsuperscript{1}, and Alex Trahan\textsuperscript{2}

\textsuperscript{1}Master of Science, Data Science, Southern Methodist University, Dallas, TX 75275 USA
\textsuperscript{2}Tandem Diabetes Care, Inc, San Diego, CA

Abstract

Millions of people live with diabetes worldwide \cite{7}. To mitigate some of the many symptoms associated with diabetes, an estimated 350,000 people in the United States rely on insulin pumps \cite{17}. For many of these people, how effectively their insulin pump performs is the difference between sleeping through the night and a life threatening emergency treatment at a hospital. Three programmed insulin pump therapy settings governing effective insulin pump function are: Basal Rate (BR), Insulin Sensitivity Factor (ISF), and Carbohydrate Ratio (ICR). For many people using insulin pumps, these therapy settings are often not correct, given their physiological needs. While existing reinforcement learning models can predict actual physiological values for these settings, they require iteration and can be slow.

The primary contribution of this research is to present a pipeline capable of providing instant predictions of close to actual patient physiological ISF, ICR, and BR from 30 days worth of data. In theory, this reduces patient waiting periods from roughly 6-8 weeks for existing reinforcement learning models to 30 days. This can serve as an aid in recommending pump therapy settings.

Data used in this study include 1,000 simulated multivariate insulin pump time series. These time series were generated by a proprietary simulator developed by Tandem Diabetes Care. This multivariate time series data also integrates simulated continuous glucose monitor (CGM) data.

This research proposes a pipeline for predicting actual patient BR, ISF, and ICR. Feature engineering, a component of this pipeline, included contextual consensus time series motif analysis. Models in the pipeline include time series native techniques such as Deep Convolutional Neural Networks (DNN) with a Long Short Term Memory input layers (LSTM) and aggregation based models such as Ridge regression and Lasso.

Aggregation based ridge regression showed the most promising results, outperforming a naive model and a DNN model. For the data evaluated and with a 20\% holdout test set, aggregate based ridge regression predicted the following normalized patient pump settings: ISF with a Mean Absolute Error of roughly 9.0\%, ICR with a Mean Absolute Error of
roughly 5% and BR with a Mean Absolute Error of roughly 6%. This is likely due to the reduction that aggregation based methods perform on each patient time series, reducing each one into a single tuple. This makes aggregation based methods less susceptible to noise and sparse signals.

One limitation in this study is that the simulated data assumes a constant value of ISF, ICR, and BR over 24 hour periods for people with diabetes. In practice, this is not the case; ISF, ICR and BR fluctuate throughout the course of a day. A future consideration would be to use simulated data with non-constant 24-hour ISF, ICR, and BR profiles.

Insulin pumps greatly improve management and outcomes for people with diabetes. Ideally, by instantly improving programmed values of ISF, ICR, and BR, people relying on insulin pumps can spend less time worrying about their pump working ineffectively, and sleep through the night knowing it is less likely they will suffer a diabetes related medical emergency. To this end, it is the hope of the researchers that the ideas, pipelines, and inference presented are further explored and tested.

1 Introduction

Millions of people worldwide suffer from diabetes mellitus [3]. This disease, when poorly managed, can be life threatening. To best mitigate the symptoms associated with diabetes, an estimated 350,000 people in the United States alone rely on insulin pumps [17]. For these people, insulin pump performance is critical to quality of life.

To manage this disease, people with diabetes have traditionally administered insulin to their blood stream multiple times per day (either manually or with an insulin pump). However, many factors affect the severity of diabetes, and quality of care can vary greatly. It can be difficult to know the correct dose of insulin needed at any given time.

Deemed by some as an “artificial pancreas,” insulin pumps in conjunction with continuous glucose monitors (CGM) can help mitigate the complexity of managing diabetes. This is accomplished by integrating with CGMs, which actively monitor patient blood sugar levels, enabling them to deliver just the right amount of insulin at the right time. Unfortunately, currently available insulin pumps and CGMs together are still a hybrid closed-loop system and require additional user-specific parameters (typically set by an endocrinologist) to operate efficiently and safely [8]. Key examples of these parameters include insulin sensitivity factor, basal rate, and insulin to carbohydrate ratio.

For those taking insulin daily, insulin to carbohydrate ratio (ICR) is meant to guide the ratio of units of insulin to available grams of carbohydrates consumed. An example ICR value would be 1:10 (for rapid acting insulin). The ICR refers to the number of available carbohydrate grams a single insulin unit will cover. Example factors affecting ICR variability are the mix of macros (carbs, protein, and fat) in a given meal and how effectively a person can uptake carbohydrates.

The rate at which a pump delivers a dose of insulin to a patient is referred to as basal rate (BR). This rate is highly variable. One cause of high BR
variability is the level of physical activity performed by a patient [19]. The standard measurement for BR is units of fast acting insulin per hour. Getting BR right is another critical component of glucose management.

How much blood glucose levels drop relative to a unit of administered insulin is referred to as the Insulin Sensitivity Factor (ISF). Standard ISF measurement is milligrams per deciliter per unit (mg/dl/unit). ISF represents patient insulin sensitivity and correlates with total daily insulin. ISF is frequently applied as a correction factor to address spikes in glucose.

Correct ISF is essential to determining short term insulin rates. For people with diabetes, it is important to get ISF right to prevent hypoglycemic events. Hypoglycemic events not only increase the probability of coma or death in extreme cases, but can impact patient day to day activities, for example lowering cognitive functioning [11].

The goal of correctly determining and setting insulin pump parameters is to minimize the frequency and duration of time spent hyperglycemic or hypoglycemic. These conditions mean that an inappropriate amount of insulin was delivered to the patient. Ideally, a patient’s blood glucose remains “in range” between the hyperglycemic and hypoglycemic thresholds. In Figure 1, an example of a blood glucose time series with hyper and hypo thresholds can be seen.

Figure 1: Glucose Thresholds as displayed on a sample blood glucose time series realization [18].

Programmed insulin pump parameters of ISF, ICR, and BR are often set or recommended by an endocrinologist to maximize time spent in range. These programmed values are intended to be as close as possible to actual physiological patient values, and are determined by factors such as past diabetes history, sex, weight, age, and recent blood glucose data. While the American Association of Clinical Endocrinology (AACE) has guidelines in place to calculate ISF, ICR,
and BR, those guidelines are for a population mean, and are used as a starting point for setting insulin pumps. There is still a significant amount of guess-and-check between patients and endocrinologists when a pump is first turned on, or when a patient undergoes a physiological change. As such, the programmed values of ISF, ICR and BR are unlikely to be close to actual for an individual patient right away.

Some algorithms exist to determine actual patient values for ISF, ICR, and BR. Many of these existing algorithms utilize forms of reinforcement learning. The power in these methods comes from how few assumptions they require. Reinforcement learning relies on optimizing for cost functions, which are derived from empirically measured values. Examples of cost functions include percentage time spent hypoglycemic and percentage of time with glucose spent in range.

Because the exploration phase can be inaccurate for reinforcement learning, one major downside is that the learning rate for this algorithm must be kept small (for example, a maximum of 10% change per day) to avoid harming patients. This means that the algorithm can in some cases take 6-12 weeks to begin to benefit patients [10], [26].

This research aims to supplement existing reinforcement learning algorithms with an algorithm that can predict close to the actual values for BR, ISF, and ICR with 30 days worth of patient data.

1.1 Prandial Context

ISF, ICR, and BR affect how blood glucose and insulin levels behave in a window of time around meals with active carbohydrates in them. People with diabetes rely on getting their pre-prandial or “pre-meal” insulin correct to avoid a spike or a crash in glucose, resulting in hyperglycemia or hypoglycemia.

Correct pre-prandial insulin level adjustment depends on whether or not a person has low or high insulin needs. For example, a person with low actual ICR needs more insulin prior to eating. If that person has a programmed value of ICR that is artificially too high, they will not receive enough insulin and may suffer from post-prandial hyperglycemia. Conversely, a person with a high actual ICR needs less insulin. If this person has a programmed value of ICR which is too low, they will likely receive too much insulin, and may experience post-prandial hypoglycemia. A pre-prandial window, where insulin is administered in advance of a meal, is typically around 30 minutes [16].

Post-prandial insulin pump data is also affected by how correct programmed values of ICR, ISF, and BR are. Post-prandial blood glucose values respond differently to insulin and consumed carbohydrates when pump settings are less correct. Because pre- and post-prandial windows correlate to pump setting accuracy, features specifically extracted from pre- and post-prandial data are important.
1.2 Contextual Consensus Motifs

Glucose time series motif analysis can be useful in detecting pre- and post-prandial patterns. These patterns can then be used as features in a model or even as a form of exploratory data analysis to determine which features to engineer.

A glucose time series motif is defined as the predominant blood glucose sub-sequence pattern in a given time series. It provides an intuitive measure of glucose sub-sequences in a CGM blood glucose time series emerge from pre-prandial and post-prandial time series as predominant patterns.

In practice, endocrinologists often view blood glucose levels and visually search for patterns that stand out in different regions of blood glucose data. Specifically for ISF and ICR, patterns present in data around meals are often analyzed the most closely. For BR, periods of fasting are often analyzed the most closely to isolate other variables. These patterns can be thought of as forms of glucose time series motifs.

Glucose time series motifs are generally determined by comparing euclidean distance between all possible (or a randomly sub-sampled set of) sub-sequence pairs. The two sub-sequences with the lowest euclidean distance between them are considered the most similar and are then considered a motif [25].

However, for a single time series realization, those sub-sequence pairs can only ever be from the same time series and are likely to exhibit patterns that are highly specific to a single patient’s blood glucose realization. This can be due to the type of environmental conditions that a particular time series was subjected to, sensor error, or even behavioral patterns that only apply directly to a single patient. Single time series glucose motifs can be considered as prone to over-fitting.

For this reason, consensus glucose motifs are used. Consensus motifs are detected by comparing sub-sequences across an array of z-normalized time series. Consensus motifs generalize better than single time series glucose motifs [12].

ISF, ICR, and BR values have complex, non-linear affects on pre-prandial and post-prandial blood glucose time series. It is likely that distinct glucose time series patterns, referred to as motifs, exist relative to how correct each programmed value is. For example, distinct pre- and post-prandial motifs may exist for the contexts where programmed values of ISF are too low vs too high; distinct pre- and post-prandial motifs may exist for contexts where programmed values of ICR are too low vs too high.

With regard to distinct motifs for contexts where programmed values of BR are too high or too low, it should be noted that fasting data is preferred over pre- and post-prandial insulin pump windows. This aligns with how BR is traditionally determined by endocrinologists in the field. Fasting data is typically used to determine basal rate because it’s affect can be more subtle and is easily masked by the affects of ISF and ICR during pre- and post-prandial windows. If programmed values of ICR and ISF are incorrect, pre- and post-prandial window patterns are dominated by that ISF and ICR discrepancy.

In order to characterize and measure occurrence frequency of these context
based patterns, consensus motif detection becomes an intriguing option. Examples of pre- and post-prandial motifs and their contexts are:

- Sharp vs Smooth Hyper and Hypoglycemic blood glucose Excursions
- Blood Glucose clipping for patients with ICR/ISF/BR that is either high, just right, or low
- Glucose spikes for patients with ICR/ISF/BR that is either high, just right, or low

Figure 2 illustrates examples of contextual hyper and hypoglycemic motifs with a sharp hyperglycemic peak, hypoglycemic clipping, and hyperglycemic wandering.

![Figure 2: Varying contextual hypoglycemic and hyperglycemic motifs. Hypoglycemic clipping can be seen near the bottom in the hypoglycemic excursions. A short hyperglycemic peak can be seen in the hyperglycemic excursions as well as hyperglycemic wandering in the top right. Note: the hyperglycemic wandering includes hyperglycemic clipping at 400 mg/dL [5].](image)

1.2.1 Novel Contributions

The goal of this research is to provide a single and instant prediction of programmed ISF, ICR, and BR that are as close as possible to the actual patient values. This results in a reduction of waiting time for programmed patient insulin pump settings to approximate actual values from roughly 6-8 weeks to 30 days for patients with new insulin pumps. Additionally, for people who have had their pumps for more than 30 days and wish to improve their performance, an immediate prediction can be made for ISF, ICR, and BR. It should
be noted that this research does not aim to compete with reinforcement learning approaches, but rather, a prediction from the models outlined in this research is intended to work in conjunction with these techniques. That is, an instant prediction can be used to approximate actual pump settings, and reinforcement learning models could then subsequently tune and continue optimizing those approximations.

This can help patients whose insulin sensitivity changes due to some medical event and can mitigate the need to “wait it out” while the pumps adjust settings iteratively. Further, an instant prediction of programmed ISF, ICR, and BR can help reduce patient dependency on endocrinologists or error prone self-research to calculate programmed pump settings. This can also serve as an aide or sanity check for endocrinologists recommending pump therapy settings. This ultimately leads to improving the healthcare practice and general care for people with diabetes using insulin pumps. With more correct values of ISF, ICR, and BR, people relying on insulin pumps can sleep through the night knowing it is less likely they will suffer a diabetes related medical emergency.

This research leverages Tandem Diabetes proprietary simulated patient insulin pump and CGM simulation time series data. By using a combination of feature engineering, contextual consensus motifs, ridge regression, and deep learning approaches like DNN with LSTM, several pipelines and resulting models predicting time-invariant insulin pump setting values are compared. The results for each of these models are then compared to a “no-information” model error which acts as a naive baseline prediction.

2 Related Works

Spanning many years, researchers across the world have studied and modeled diabetes. One important focus of this research has been on short term forecasting of blood glucose, specifically for 30 to 60 minutes. Although the goal of short term glucose forecasting differs from the goal of the research in this paper, there are also some similarities.

2.1 Time Series

Because blood glucose data often comes in a time series format, autoregressive models have traditionally been used to predict glucose levels 10, 20 and 30 minutes into the future [6]. In their research, Frandes, Tomar and Lungeanu attempted to predict glucose levels by leveraging deterministic chaotic properties of glucose dynamics [1]. By using these properties as model features, the autoregressive model outperformed previous autoregressive models [6]. Although this research will not continue the exploration of chaotic properties on time series or classical time series models, it will explore statistical methodologies to aggregate time series into a single tuple similarly to [4]. The main difference with [4] will be the objective of the research and the features used for aggregation.
2.2 Reinforcement Learning

Many proposed or existing algorithms exist which implement reinforcement learning to predict various features of insulin pumps. In their research, Jfar, El Fathi and Haidar propose an algorithm for iteratively updating ICR and BR by leveraging reinforcement learning. In particular, a technique called Q-Value learning is used [10]. After approximately 5 weeks of training, their model was able to improve the percentage of time spent in range for people with insulin pumps from 67% to 86% of the time.

Often times, a type of deep reinforcement Q-Learning algorithm called Soft Actor Critic (SAC) is used to improve some of the classical pitfalls of Q-value reinforcement learning algorithms. In their research, “A Blood Glucose Control Framework Based on Reinforcement Learning With Safety and Interpretability” [15], the authors implemented a SAC model with Adaptive safe actor models. This was done to prevent the reinforcement SAC models from exploring values for a patient known to be unsafe, and was accomplished through an actuator of safety which pre-defines actions and rules to rescue the patient from hypoglycemia or hyperglycemia.

Although these models are effective, and may seem appealing, they come with a caveat. They rely on patients spending up to 6 to 12 weeks in an exploratory phase. Accordingly, during this period, pump settings are not optimal. This research intends to work in conjunction with existing iterative or reinforcement learning techniques, to get the values of ISF, ICR, or BR close to actual values in a shorter period of time resulting in fewer unwanted fluctuations in settings. This research will not continue exploration in reinforcement learning, but rather explore an outright prediction to initialize a pump with approximated values, where an iterative or reinforcement based technique could then be used to further fine tune pump settings.

2.3 Neural Networks:

Artificial neural networks are used to solve a wide variety of machine learning problems, including medical diagnosis and treatments. While there are a vast number of types of neural networks, two frequently used choices for prediction when considering sequential multivariate time series diabetes data are Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) [24]. LSTM and CNN are traditionally purposed for different problems. In the case where the goal is to forecast or predict sequential data like blood glucose time series, LSTM’s are very popular. However, this study aims to predict static numerical values like ICR, ISF, and BR in a non-causal manner from 30 days worth of insulin pump and CGM data. To meet these requirements, 1-Dimensional multivariate Deep Convolutional Neural Networks can be leveraged due to their innate ability to detect temporal patterns across inputs using an entire 1-Dimensional multivariate timeseries [13].
2.3.1 Deep Convolutional Neural Networks

Deep Convolutional Neural Networks (DNNs) generally make use of multiple connected latent or “hidden” layers of convolution, pooling, and activation. As proposed in [23], dropout can also be added to a deep CNN to reduce over-fitting via regularization.

In their work predicting blood glucose in [14], the researchers leveraged multivariate DNNs for feature extraction. These features were subsequently used by LSTMs to predict blood glucose. Latent feature maps were used as inputs. A key difference between their research and this research is the proposed model in this research is meant to generalize to any patient [14]. However, the use of latent features with LSTMs in combination with DNNs is intriguing and is of interest in this study.

In their benchmarking of various deep neural network algorithms across a variety of time series problems [9], demonstrated that DNNs for multivariate time series data ranked well among other existing state of the art TCN models. This research will continue an exploration of DNN models as applied to insulin pump data, with heavy emphasis on feature engineering including use of contextual consensus motif analysis.

2.4 Motif Analysis

In their research with CGM data, [5] proposed contextual motifs. The researchers used contextual motifs to increase hypo- and hyper-glycemic event prediction when compared to the naive motif detection without context. For example, a context might be as simple as recording the size of a meal in carbohydrates. Further, in [5] the researchers proposed a generative model using joint inference to infer each context via a probability distribution of motifs. Their joint inference model performed well when compared to a traditional motif mixture model (MMM). The primary inspiration in the research [5] as applied to this study is that when context is available for motif detection, it can be leveraged.

Further mitigating pitfalls of motif analysis, the researchers in [12] propose consensus motif detection. Rather than searching for a single motif in a single time series, motifs are detected across multiple time series. This assumes those time series are all scaled the same and result from the same process. As is highlighted, consensus motif analysis is generally preferred due to its broad applicability and innate ability to counter the effects of over-fitting by finding motifs in a single time series.

2.5 Glycemic Control & Correlated Features

CGM glucose area under the curve is a useful post-meal or post-prandial metric. In their research in [22], using blood glucose estimates from CGM and minimally invasive interstitial fluid extraction technology (MIET), area under the curve for up to 8 hours after meals is explored. Additionally, in [20], post-prandial hyperglycemia AUC was compared to different blood glucose measurement techniques.
like MIET to determine significance in correlation between non-blood sampling based glucose estimation techniques and established CGM references.

In [21] Insulin to Carbohydrate Ratio (ICR) optimization is explored with respect to pre-meal boluses in order to minimize the number of hypoglycemic events patients experience. The research shows that that optimizing ICR can significantly reduce the number of hypoglycemic events. As this correlation exists, it is likely valuable to include features which capture the number of hypoglycemic events a patient experiences in an attempt to predict, at the very least, ICR.

Highlighted in [2], where insulin sensitivity was assessed in three groups by the euglycemic-hyperinsulinemic clamps, post-prandial hyperglycemia is correlated with insulin sensitivity differently for certain ethnicities. A post-prandial peak, or glucose area under the curve is used to capture this correlation. As such, it makes sense to build features which capture post-prandial hyperglycemia when trying to predict ISF.

Motivated by prior research and analysis, this work demonstrates that combining context driven feature extraction via motif analysis combined with Deep CNN techniques on multivariate CGM and insulin pump data can reduce the amount of time people with insulin pumps spend in a guess-and-check loop, optimizing their ICR, ISF, and BR settings.

These related works have both informed and inspired this research to predict ISF, ICR, and BR. While it is not the goal of this study to implement a reinforcement learning model, the models proposed in this research are intended to work in conjunction with existing algorithms.

3 Methods

This research details methods and models used to build predictive models and pre-processing pipelines for ISF, ICR, and BR. This includes the proprietary Tandem Diabetes Care simulation data used in the study, the study’s dependent target variables of ISF, ICR, and BR insulin pump setting ratios, pre-processing performed on the data (such as low pass filtering), simple feature engineering (such as adding a ‘meal status’ window), complex feature engineering (such as detecting contextual motifs), testing strategies, and predictive models used in the study.

3.1 Data

The data used for this research is generated using a proprietary Tandem Diabetes Care simulator. It statistically emulates real data from type 1 diabetes patients. This proprietary data generation is performed using statistical distributions obtained from real patient data to recreate realistic physiological and CGM patterns along with insulin pump input data and responses.

Each simulated insulin pump’s data is represented by a multivariate time series with 8,692 time steps. Each time step represents a simulated measurement
obtained at an interval of 5 minutes. This equates to approximately 30 days worth of data. In this study, 1,000 insulin pumps were simulated.

Each multivariate time series consists of features that emulate data real patients with insulin pumps would have when using a Tandem pump. These features include CGM estimates of Blood Glucose, 30-minute forecasts of blood glucose levels, consumed carbohydrates announced, units of insulin suggested, programmed pump settings of ISF, ICR, BR, and a running daily insulin estimate. For the purposes of this study, the hypoglycemic threshold is considered 70 mg/dL and the threshold for hyperglycemia is 180 mg/dL.

In particular, programmed pump settings of ISF, ICR, and BR represent features which are always off by some amount when compared to their actual values for any given patient. These values are traditionally set and subsequently tuned with the direction of an endocrinologist and guidelines set forth by the American Association of Clinical Endocrinology (AACE).

In addition to the programmed values, each time series is associated with the actual settings for each patient. These actual values are time series simulation hyperparameters. These simulation hyperparameters include actual ISF, actual ICR, and actual BR. These actual values are used in the proprietary simulation to generate data and represent actual physiological response to carbohydrates and insulin. Outside of simulation a patient cannot measure these directly.

The goal of this research is to predict how incorrect programmed ICR, BR, and ISF are when compared to actual patient values by examining dependent variables like blood glucose, estimated daily insulin received, carbohydrates consumed, programmed ICR, BR, and ISF, glucose predictions 30 minutes out, and so on. This is accomplished by modeling or otherwise capturing potentially complex and non-linear temporal relationships and interactions that emerge in the simulated CGM and insulin pump data.

Features included in the simulation data used in this research are:

<table>
<thead>
<tr>
<th>Dependent Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbohydrate grams announced (grams, numerical)</td>
</tr>
<tr>
<td>Glucose prediction for the next 30 minutes (mg/dL)</td>
</tr>
<tr>
<td>CGM glucose estimate (mg/dL)</td>
</tr>
<tr>
<td>Units of fast-acting insulin suggested (units)</td>
</tr>
<tr>
<td>Manual bolus commanded (units of insulin)</td>
</tr>
<tr>
<td>Total estimated Daily Insulin dose (units of insulin /day)</td>
</tr>
<tr>
<td>Hypoglycemic event (was a hypoglycemic event detected, categorical)</td>
</tr>
<tr>
<td>Programmed Insulin Sensitivity Factor (ISF) (mg/dL/unit of insulin)</td>
</tr>
<tr>
<td>Programmed Basal Rate (BR) (units/hr)</td>
</tr>
<tr>
<td>Programmed Insulin to Carbohydrate Ratio (ICR) (grams of carbohydrate/unit of insulin)</td>
</tr>
</tbody>
</table>

Table 1: Dependent variables from the multivariate insulin pump data used in this study.
### Independent Variables

<table>
<thead>
<tr>
<th>Actual Insulin Sensitivity Factor (ISF) (mg/dL/unit)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Basal rate (BR) (Units/hr)</td>
</tr>
<tr>
<td>Actual Insulin to Carbohydrate Ratio (ICR)</td>
</tr>
<tr>
<td>(grams of carbohydrate/unit of insulin)</td>
</tr>
</tbody>
</table>

Table 2: Independent variables which are simulation hyperparameters used to generate the multivariate insulin pump data for this study.

### Target Variables

In order to standardize target variables for the data to represent how incorrect programmed vs actual ISF, ICR and BR across various patients can be, ratios are used. This standardization via a ratio is necessary for generalization. For example, a recommended programmed ICR change of 1 is bigger for someone with a 10:1 ratio when compared with someone at 14:1.

The same approach of a programmed vs actual ratio is used for ISF and BR. Figure 3 displays the distributions for each of the target variables used in this study. The somewhat log normal distribution of each is apparent.

![ISF Distribution](image1)

(a) ISF Distribution

![BR Distribution](image2)

(b) BR Distribution

![ICR Distribution](image3)

(c) ICR Distribution

Figure 3: Distribution of Target Variable. Distribution (a) shows the normalized ISF target variable. Distribution (b) shows the normalized BR target variable. Distribution (c) shows the normalized ICR target variable.
**Feature engineering**

Several additional explanatory variables were created to capture more useful information from each patient time series. These features were created to approximate what endocrinologists focus on when setting pump values, de-noise the time series data, and highlight patterns in the data.

In order to normalize variables in each of the patient time series, all of these variables were scaled and centered. A new log transformed version for all continuous explanatory variables was also added before scaling the data.

For all of the continuous explanatory variables, rolling window averages and standard deviations were created. For this, a window size of 10 minutes was used.

Insulin pump data gathered from the short window of time ranging from just before a meal with carbohydrates in it to several hours after eating a meal with carbohydrates in it characterizes diabetes and specifically ICR and also ISF.

For this reason, a custom categorical feature called “meal status” was implemented. This “meal status” marks time-steps categorically as “pre-meal” (pre-prandial) when they are 30 minutes or less directly before a time-step with at least 5 grams reported carbohydrates consumed, “meal” which is the actual time-step where at least 5 grams of consumed carbohydrates are reported, and post-meal (post-prandial) which is up to 200 minutes after a meal with at least 5 grams of carbohydrates in it. Note: 3.3 hours was a useful amount of time to choose, because it was represented by 40 time-steps. It is similar to a 4 hour post meal glucose AUC and represents a form of blood glucose momentum.

Due to the potential for meals to be near to one another in time, each categorical meal status variable is written in an exclusive manner with the following order of priority 1. “post-meal” 2. “meal” 3. “pre-meal” and 4. “fasting.”

The following diagram shows meal status based glucose values
Figure 4: These kernel density plots represent the distributions of glucose values contained within each meal status context. Note: post-meal glucose has a wider range and higher peak than other contexts.

**Aggregation Features**

In order to feed time series data into aggregation based regression models, the original time series data was aggregated such that each time series is reduced into a single tuple. This reduction transforms the existing data from an array of multivariate time series to a single table of 1000 rows (1-per patient). Aggregate statistical values calculated for each feature included the median, mean standard deviation, minimum, maximum, geometric mean, and geometric standard deviation.

Additional aggregate features were added including:
## Aggregate Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percent time clipping</td>
<td>Percentage time blood glucose is clipping at 400 mg/dL</td>
</tr>
<tr>
<td>Percent time hypoglycemic</td>
<td>Percentage time a patient’s glucose spends below 54 mg/dL</td>
</tr>
<tr>
<td>Percent time above range</td>
<td>Percentage time a patient’s glucose spends below 70 mg/dL</td>
</tr>
<tr>
<td>Percent time hyperglycemic</td>
<td>Percentage time a patient’s glucose spends above 140 mg/dL</td>
</tr>
<tr>
<td>Percent time in range</td>
<td>Percentage time a patient’s glucose spends above 250 mg/dL</td>
</tr>
<tr>
<td>Percent time in tight range</td>
<td>Time a patient’s glucose spends between 70 mg/dL and 180 mg/dL</td>
</tr>
<tr>
<td>Post-prandial glycemic peak</td>
<td>Time a patient’s glucose spends between 70 mg/dL and 140 mg/dL</td>
</tr>
<tr>
<td>200 minute post meal glucose AUC</td>
<td>Mean glycemic peak within 200 minutes of a mean</td>
</tr>
</tbody>
</table>

Table 3: Additional Aggregate Feature

### Contextual Consensus Motifs

In this research, contextual consensus motifs features are leveraged as both aggregate and time series native features. Non-aggregate motif based features include contextual motifs, located in each time series. Aggregation motif based features include contextual motif counts and ratios. Context boundaries are determined by segregating glucose data for patients by context. Specifically, 9 various contexts are explored for motifs of size 15 time steps (75 minutes) and also 20 time steps (100 minutes):

<table>
<thead>
<tr>
<th>Features</th>
<th>Predicate</th>
</tr>
</thead>
<tbody>
<tr>
<td>low ISF meal motif</td>
<td>ISF target variable less than 0.75</td>
</tr>
<tr>
<td>good ISF meal motif</td>
<td>ISF target variable greater 0.95 than AND less than 1.05</td>
</tr>
<tr>
<td>high ISF meal motif</td>
<td>ISF target variable greater than 1.25</td>
</tr>
<tr>
<td>low ICR meal motif</td>
<td>ICR target variable less than 0.75</td>
</tr>
<tr>
<td>good ICR meal motif</td>
<td>ICR target variable greater 0.95 than AND less than 1.05</td>
</tr>
<tr>
<td>high ICR meal motif</td>
<td>ICR target variable greater than 1.25</td>
</tr>
<tr>
<td>low BR fasting motif</td>
<td>BR target variable less than 0.75</td>
</tr>
<tr>
<td>good BR fasting motif</td>
<td>BR target variable greater 0.95 than AND less than 1.05</td>
</tr>
<tr>
<td>high BR fasting motif</td>
<td>BR target variable greater than 1.25</td>
</tr>
</tbody>
</table>

Table 4: Contextual Consensus Motifs

In order to determine contextual “seed” motifs for each of the above contexts, estimated CGM blood glucose for each context is first parsed from the original time series. Meals are identified by reported carbohydrates consumed, and a
sub-sequence of that time series is windowed around that meal (pre-prandial 30 minutes to post-prandial 200 minutes). In the case of overlapping meals that are close in time, post-prandial labeling for an ongoing meal is preferred over the pre-prandial labeling for the next meal. Sub-sequence windows around a meal are labeled “meal” where those in between are labeled “fasting” sequences. Some filtering is then performed to ensure that the sub-sequences used are long enough to perform motif detection on them (a minimum of 3x the motif window size was arbitrarily chosen).

The resulting lists of contextual CGM blood glucose sub-sequences are then independently run through an Ostinato algorithm and z-normalized sub-sequence euclidean distance is compared for each sub-sequence. After estimated blood glucose sub-sequence motifs are determined by finding the sub-sequence pair with the lowest distance between them, these contextual consensus “seed” motifs are used to generate a latent motif distance for each training patient blood glucose time series. This new latent time series feature represents how similar, at any given time step, the training glucose data is to each seed motif. This is achieved by convolving the seed motifs with the z-normalized training glucose time series data for each patient.

Next, the lowest distance for each context for these latent motif distance time series are indexed and sorted in a ranked list. For example, the lowest “good ISF meal motif” distance is ranked, with a single entry in the ranked list representing a single meal sequence. The smallest 4% of the ranked distances per context are considered as being an occurrence of that particular motif. The ranked motif distance at the arbitrary cutoff threshold of 4% is then stored as the cutoff threshold, to be used to determine whether or not a contextual motif is located at a particular time-step in new data.

Finally, for all training and test blood glucose estimates, the seed motifs are convolved and the similarity distance latent time series is again generated. For each context, the cutoff thresholds determined previously are used to determine if a motif was detected at a particular time-step. This calculates where each motif appears in each time series. At this point, the location based features are added to the multivariate time series data.

Examples of motifs discovered with this technique can be seen in Figure 5, overlaid for 40 different patients:
As can be seen in Figure 5, the z-normalized contextual glucose motifs for low BR and low ICR are visually different to good and high BR and ICR. There is also a higher variance for the Low BR and ICR motifs.
Figure 6: Contextual Motif Detection on glucose time series for 2 patients. Motifs are represented by overlaid vertical columns on each time series. Note: CR is an alias for ICR.

In Figure 6, detected context based motifs as displayed for patient #0 and patient #1. The context based motifs include: low BR, low ICR, low ISF, good BR, good ICR, good ISF, high BR, high ICR, and high ISF. It can be noted that certain motifs have a higher prevalence across each glucose time series.

For Aggregate models like Ridge Regression, simple counts of each contextual motif occurrence per patient time series are performed, along with determining
the contextual motif count ratios (normalized by how many motif occurrences were detected for all ISF, BR, or ICR contexts). These counts and ratios are then added as aggregation based features used in aggregation based models. The correlation of these aggregate features with our ISF, ICR, and BR target variables can be seen in Figure 7:

![Figure 7: Aggregate Features generated from motif detection vs ICR, ISF, and BR target variables (with overlaid general linear regression)](image_url)

Test - Train Split.
In order to evaluate the proposed models without introducing bias or suffering from data leakage, the data was randomly split into train and test sets. The split used was: 80% of the data was used for training and 20% of the data was used for testing.
was used for testing.

### 3.2 Models

Nine different models were explored in order to compare their prediction accuracy and to compare inference. Out of the nine models, eight were supervised predictive algorithms and one was a naive model, used as a baseline. These models either fit into a category of aggregation based models or time series native based models.

#### 3.2.1 No information

In order to represent the strongest naive prediction, a “no information” model is created. This model simply predicts patients’ programmed ISF, ICR, and BR values to have no change. The mean absolute error is then the mean absolute difference between 1 (a perfect ratio of actual to programmed settings) and the
calculated actual vs. programmed setting ratio (representing the error associated with a prediction of no change). The mean absolute errors for each ratio are highlighted in equations [1], [2] and [3] is then:

$$MEA_{NoInformationISF} = \frac{\sum_{i=1}^{n} |(1 - \frac{ISF_{Actual}}{ISF_{Programmed}})|}{n}$$ (1)

$$MEA_{NoInformationICR} = \frac{\sum_{i=1}^{n} |(1 - \frac{ICR_{Actual}}{ICR_{Programmed}})|}{n}$$ (2)

$$MEA_{NoInformationBR} = \frac{\sum_{i=1}^{n} |(1 - \frac{BR_{Actual}}{BR_{Programmed}})|}{n}$$ (3)

Where $n$ is the number of patients sampled, MAE stands for mean absolute error.

### 3.2.2 Time series native Models

In order to detect patterns and evaluate the efficacy of a multi-output model which uses patient pump time series data in original form, non-aggregation based models and their associated pre-processing pipelines are evaluated. These sub-pipelines are referred to as “time series native”.

**Low Pass Filtering**

In an effort to compensate for sparse events and potentially noisy time series, low pass filtering via a Fast Fourier Transform (FFT) is implemented. This low pass filter is implemented as a pre-processing layer upstream to the input to the time series native models, including a DNN. This method helps in de-noising the data by first transforming it into the frequency domain. Once the data is represented in the frequency domain, a low pass filter passes only the bottom 1% of the lowest frequency bins. Figure 9 displays an example of how the FFT Low Pass Filter de-noises the time series data set for patient 0.
Figure 9: Fast Fourier Transform Low Pass Filtering for Patient #0

DNN
The low pass filtered time series data is then fed as an input layer of a 1-dimensional deep convolutional neural network (DNN) capable of predicting the ISF, ICR, and BR corrections from the simulated multivariate time series insulin pump data. In order to support multi-output, this DNN model was built using the Keras Functional API. This API allows greater customization when architecting the design of the model and supports further expansion in the future.

As seen in Figure 10, the DNN is divided into three different branches. Each branch contains the same hidden layer architecture but is trained independently. The hidden layer architecture consists of eight different blocks. The first block consists of a standard feed forward LTSM model. The second block consists of a 1D convolutional layer with eight filters followed by an activation function. The next six blocks consist of a 1D max-pooling layer, followed by a drop-out layer, a 1D convolution, and a Relu activation transform.
Long Short Term Memory Layer

In order to capture multivariate interactions over spans of time, a Long Short Term Memory (LSTM) layer is implemented in conjunction with the deep CNN. The primary purpose of this layer is to predict a latent embedding value which is then used as an input to the DNN. An LSTM is used here because of its ability to “remember” a weighted percentage of context from previous time-steps. In the case of this research, the LSTM layer may help build in memory of pre-prandial insulin dose and also blood glucose levels when predicting context for a post-prandial time-step.

Convolutional layers

The proposed DNN model uses a 1D convolutional layer which convolves in the time series steps for each of the independent variables. These convolutions help capture temporal entropy out of the time series data. The model contains a total of six 1D convolution layers with filters values ranging in size from 8 to 512.

Max Pooling Layers

Each DNN block also includes a 1D max-pooling layer which helps to extract the most important features from the feature map created by the convolution layer preceding it. For each of these layers, pool size values ranged from three to nine.
The final layers for each branch consist of the following layer sequence: flat-
tening, followed by a fully-connected layer with 512 units, followed by a Relu
activation and a batch normalization layer. Additionally, a drop-out layer is
added, followed by a dense layer containing only one unit and a Relu activation
function.

For the DNN model, hyperparameters included 500 epochs, early stopping
with a patience of 30 batches, and monitoring by validation loss.

3.2.3 Aggregation Based Models

In order to detect patterns and evaluate the efficacy of a multi-output model
which aggregates each patient pump time series into a single tuple, aggrega-
tion based models and their associated pre-processing pipelines were evaluated.
These models regress on aggregated feature data of 800 training entries with a
holdout set of 200 entries for test evaluation.

3.2.4 Regression

Seven different aggregation based regression models were trained, including
Ridge, Random Forrest Regression, Ridge with Recursive Feature Elimination,
and LASSO. To choose the best hyper-parameters for each model, grid searches
were performed.

The search space for the Ridge Regression with and without Recursive fea-
ture elimination consisted of 56 combinations each. For both models, the com-
bination with the best performance consisted of the following hyperparameters:

- L2 regularization strength, ALPHA = 10
- Solver: Cholesky
- Fit intercept: True
The search space for the Lasso Regression with Recursive feature elimination consisted of 20 combinations. The combination with the best performance consisted of the following hyperparameters:

- L1 regularization strength, ALPHA = 0.01
- Model selection: Random
- Fit intercept: True

The search space for the Random Forrest regression consisted of 36 combinations. The combination with the best performance consisted of the following hyper-parameters:

- Criterion: Squared error
- Max Features: Auto, which means that the algorithm automatically chooses the strategy to select the number of max features based on performance.
- Number of Estimators: 5,000

The search space for the Random Forrest regression using recursive feature elimination consisted of 36 combinations. The combination with the best performance consisted of the following hyper-parameters:

- Criterion: Squared error
- Max Features: Auto, which means that the algorithm automatically chooses the strategy to select the number of max features based on performance.
- Number of Estimators: 2,000

The search space for the Ridge Regression using the features selected from the Lasso model consisted of 56 combinations. The combination with the best performance consisted of the following hyper-parameters:

- L2 regularization strength, ALPHA = 10
- Solver: Least Squares, which is the fastest solver using the regularized least squares procedure.
- Fit intercept: True

The search space for the Random Forrest regression using the features selected from the Lasso model consisted of 36 combinations. The combination with the best performance consisted of the following hyper-parameters:

- Criterion: Squared error
- Max Features: Auto, which means that the algorithm automatically chooses the strategy to select the number of max features based on performance.
- Number of Estimators: 2,000
4 Results

For this research, the aggregation based, time series native, and No-Information predictive models are compared. The evaluated models were trained using proprietary Tandem Diabetes simulated CGM and insulin pump data from 30 days for 1000 patients. In both models the chosen validation metric was the mean absolute error of the predicted normalized target settings. Because this error is normalized proportional to the values of programmed settings, it represents the percentage of mean absolute error.

The proposed model with the lowest mean absolute error is consistently Ridge Regression. When predicting pump setting target variables, it outperforms the no information model by 47% for ISF, 72% for ICR, and 71% for BR. Compared to the DNN the Ridge regression outperforms it by 55% for ISF, 75% for ICR, and 70% for BR.

Mean absolute error when predicting ISF, ICR, and BR target variables for each model can be seen in the Table 5:

<table>
<thead>
<tr>
<th>Model</th>
<th>ISF</th>
<th>ICR</th>
<th>BR</th>
</tr>
</thead>
<tbody>
<tr>
<td>No Information Error</td>
<td>0.17</td>
<td>0.17</td>
<td>0.21</td>
</tr>
<tr>
<td>DNN</td>
<td>0.20</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>Random Forrest Regression</td>
<td>0.10</td>
<td>0.06</td>
<td>0.05</td>
</tr>
<tr>
<td>Lasso Regression</td>
<td>0.12</td>
<td>0.09</td>
<td>0.07</td>
</tr>
<tr>
<td>Ridge Regression with Recursive Feature Elimination</td>
<td>0.09</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>Random Forrest Regression with Recursive Feature Elimination</td>
<td>0.14</td>
<td>0.11</td>
<td>0.08</td>
</tr>
<tr>
<td>Ridge Regression using features from Lasso</td>
<td>0.11</td>
<td>0.08</td>
<td>0.10</td>
</tr>
<tr>
<td>Random Forrest Regression using features from Lasso</td>
<td>0.14</td>
<td>0.11</td>
<td>0.11</td>
</tr>
</tbody>
</table>

Table 5: Models Result Table (Mean Absolute Error)
Figure 12 displays the train and test mean absolute error per epoch. Note: due to the early stopping criteria, the model ultimately only trained for 100 epochs.

4.1 Feature Importance

One important benefit of the aggregation based Ridge regression model is that comparing feature importance can be done both by extracting feature coefficients and using Shapley additive values for each feature. Coefficient based feature importances make Ridge Regression easier to interpret, while Shapley values can be generated for all model types, providing a good comparison between models. Figures 13 - 16 display both Shapley and coefficient feature importance for each of the target variables.
Figure 13: Coefficient and Shapley based feature importance for ISF
Figure 14: Coefficient and Shapley based feature importance for BR
Figure 15: Coefficient and Shapley based feature importance for ICR
Figure 16: Aggregate Coefficient and Shapley based feature importance
5 Discussion

There are several distinct reasons why the models compared in this study performed differently. Additionally, it is important to consider why motif detection results show certain dominant motifs reappearing across multiple contexts.

5.1 Motif Features

Interestingly, it appears that there are dominant motifs across many contexts. One such motif was hyperglycemic clipping. This appeared when the window size was increased for motif detection. This clipping was the result of glucose values being so high that the maximum possible glucose reading of a CGM was exceeded. It presented as a dominant signal for several contexts. This is likely because increasing motif size favors more specific or less generalized motifs that are then dominated by fewer patients with very similar to exactly the same clipping behavior.

Another motif which appears in many contexts is a peak in glucose. This is likely because glucose peaking followed by decreasing in value should happen after every meal. It was the goal of this study to at least encourage motif analysis to find subtle differences between the motif peaks, but in general many look very similar. This leads to an unexpected correlation for the various contextual motif counts. Lastly, the most correlated motif feature appears to be 'Good Insulin to Carbohydrate Ratio Meal Motif Count'.

When assessing feature importance for Ridge Regression, one can see that the motif based features do not ultimately end up in the top three most important Ridge features. However, when a deeper feature importance is considered, it can be seen that some of the motif count features do carry significance in the model. For example, ISF Meal Motif count was the 14th most important feature in predicting ISF. There may be high redundancy between motif features and other features that were engineered leading to high multi-collinearity. This likely means that motif analysis and its associated feature generation is a useful tool for both highlighting and building other features by hand as part of the pipeline highlighted in this research.

5.2 DNN

Surprisingly, the DNN model tested in this research did not perform well in terms of prediction mean absolute test error. In fact, DNN performed worse than the ridge regression model and was largely similar in prediction accuracy to the naive no-information model with a test mean absolute error of 0.20 for the ISF, ICR and BR target variables. This was unexpected because intuitively, the DNN should be able to capture more complex spatial and temporal patterns that emerge in time series data. It is likely that the sparsity of certain signals in the time series leads to increased noise, a DNN will generally attempt to over-fit.
Another problem with the DNN model is that in spite of a self-attention model and Shapley scores, the model is complex and difficult to interpret. For these reasons, and in the context of an algorithm which needs to be tested or eventually FDA approved, complex models such as these are generally not preferred. Although feature importance could be implemented on the DNN, it performs similarly to the no-information model and worse than Ridge regression. It does not make sense to perform feature importance analysis or implement a self-attention layer on the DNN model.

5.3 Ridge Regression

Ridge regression performed comparatively well at predicting pump settings. It is likely this is in part due to the aggregation and signal compression that takes place on every multivariate time series, reducing that time series to a single tuple. The aggregation means that the sparse impulse like signals are captured well and over-fitting noise is not as much of a factor.

Interestingly, for most of the evaluated models, predicting ISF appears to be more difficult than predicting ICR or BR. For ridge regression, this appears to be no different. That said, Ridge regression still beats the naive No-Information model for all three target variables.

When analyzing the Ridge regression feature importance, there are two primary metrics compared. One is the scaled regression coefficient based importance, and the other involves using Shapley scores.

Additionally, for the top five features in predicting ISF, ICR, and BR, multicollinearity was assessed for the aggregate features used in Ridge Regression. In particular, three features were highly correlated with other features:

1. 'Rolling window standard deviation of estimated blood glucose'
2. 'Rolling window median value of estimated blood glucose'
3. 'Rolling window average of the time differenced blood glucose standard deviation'

Of these three highly multi-collinear features, when 'Rolling window standard deviation of estimated blood glucose' and 'Rolling window median value of estimated blood glucose' were removed, Ridge regression performed with very similar prediction accuracy. However, when 'Rolling window average of the time differenced blood glucose standard deviation' was removed, test prediction accuracy for ISF dropped by roughly 1%.

Consequently, 'Rolling window standard deviation of estimated blood glucose' and 'Rolling window median value of estimated blood glucose' were removed from the ridge regression model, and 'Rolling windows average of the time differenced blood glucose standard deviation’ is left in the model.

After removing the features listed above, correlation heat maps can be seen in figures 17 - 19. It can be noted that there is significant correlation between remaining features.
Figure 17: Correlation Heatmap for ISF

Figure 18: Correlation Heatmap for BR
With regard to coefficient based feature importance, the three most important features when predicting ISF (in order) are:

1. ISF programmed base
2. Units of insulin suggested max
3. Mean post-prandial glycemic Peak

With regard to Shapley based feature importance, the three most important features (in order) are:

1. Units of insulin suggested max
2. ISF programmed base
3. Mean post-prandial glycemic peak

While the order of coefficient and Shapley based feature importance's are slightly different, as seen in the feature importance plots in Figure 13, the top three features are all the same.

The importance of “ISF programmed base” is expected, as this is the starting ISF value that patients have their pumps set to. As evidenced by the No-Information prediction, most patients have their ISF values programmed to within 17% of the correct value before any prediction is made by one of the above models. It should be noted that “Units of insulin suggested” features are actually summary statistics on variables which come from another proprietary
model which takes into account ICR, ISF, and BR. It is likely that these features not only match the no information rate of “programmed settings,” but pick up other predictive entropy or information that the proprietary model also captures.

**BR**

With regard to coefficient based feature importance, the three most important features when predicting BR (in order) are:

1. Mean rolling window standard deviation of blood glucose
2. Percentage time clipping
3. Units of insulin suggested standard deviation

With regard to Shapley based feature importance, the three most important features (in order) are:

1. Mean rolling window standard deviation of blood glucose
2. Percentage time clipping
3. Units of insulin suggested standard deviation

In the case of BR, the coefficient and Shapley based feature importances are slightly different, although the top 10 features largely agree. See Figure 14.

Interestingly, Percentage time clipping is found to be highly important to both Shapley and Coefficient based feature importance. This value represents the percentage of time a patient spends at or above the upper limit of the CGM used in this study (400 mg/dL). An interesting note is that the clipping feature was initially discovered through motif analysis.

One feature, called the mean rolling window standard deviation of blood glucose, represents how much estimated average blood glucose variance there is for one patient. Simply put, how much a patient’s blood glucose swings. Possible reasons for correlation between this average glucose swing for a patient and BR could be, when a BR value is too high, blood glucose crashes easily (having potential side effects for patients). Conversely, when BR value is too low, blood glucose might stay high for too long.

Units of insulin suggested standard deviation is also highly important for both Shapley and coefficient based values. This can be interpreted as “how much work the insulin suggestion decision support algorithm is doing”. Theoretically, a higher standard deviation for this feature means that the insulin suggestion algorithm is fluctuating heavily and it is likely that the programmed BR value is far away from the actual patient value.

**ICR**

With regard to coefficient based feature importance, the three most important features for ICR (in order) are:

1. Mean 200 minute post-meal glucose area under curve
2. ICR programmed settings
3. Percent time clipping

With regard to Shapley based feature importance, the three most important features (in order) are:

1. Mean 200 minute post-meal glucose area under curve
2. ICR programmed settings
3. Rolling window standard deviation for the standard deviation of time difference blood glucose

While the order of coefficient and Shapley based feature importances are slightly different, like the other features and as seen in the feature importance plots in Figure 15, the top two features are the same.

Mean 200 minute post-meal glucose area under curve is significantly more important than the next most important features for both Shapley and Coefficient based measures. A high value means that glucose values stayed high for too long after a patient's meal. When glucose values stay high for too long, negative consequences for patient health can arise. Therefore, adjusting ICR to lower 200 minute post meal blood glucose area under the curve may be a priority for endocrinologists and patients alike.

**Overall Feature Importance**

With regard to overall coefficient based feature importance, the three most important features (in order) are:

1. Mean rolling window average of blood glucose mean (log transformed)
2. Minimum number of hypoglycemic events
3. Mean 200 minute post-meal glucose area under curve

With regard to Shapley based feature importance, the three most important features (in order) are:

1. Mean 200 minute post-meal glucose area under curve
2. Median rolling window average of standard deviation of time differenced blood glucose
3. Standard deviation of rolling window standard deviation of blood glucose

It should be noted that the overall feature importance for Coefficient and Shapley based feature importance agree less than they do for the individual target variable based feature importance. These overall importance measures are meant to provide insight, only in the case where all predicted pump settings (ISF, ICR, and BR) are considered equally important. This may be unlikely in
practice. They are determined by simply stacking the importances from ISF, ICR, and BR Ridge regression feature importance.

In this case, mean 200 minute post-meal glucose area under curve and mean rolling window average of blood glucose mean (log transformed) emerge as the top features from feature importance metrics. Intuitively similar to ICR, the amount of time a patient’s blood glucose remains high (or low) after a meal is largely correlated to insulin pump settings. Mean rolling window average of blood glucose mean (log transformed) captures the average blood glucose value (after distribution correcting via a log transform) and demonstrates that the average blood glucose value of a patient will correlate highly to their pump settings.

5.3.1 Limitations in this study

It should be noted that this study used 1,000 simulated patients. These patients were simulated with constant ICR, ISF, and BR data. In practice, real patient ICR, ISF and BR fluctuates throughout the course of a day. The results from this study are meant to generalize to the mean monthly ICR, ISF, and BR. However, to test this thoroughly, data with fluctuating 24 hour profiles for ICR, ISF, and BR may be required.

This research is also complicated by the nature and underlying structure of the simulated, multivariate multi-time series data. Many existing tools or libraries for time series are not built to solve this sort of problem without modification.

Another challenge of this work is that the no-information model performs reasonably well. This makes sense as patients generally have their programmed ISF, ICR, and BR pump settings reasonably close to their actual physiological values. Patient’s lives depend on it. If someone using an insulin pump has ISF, ICR, or BR which is severely incorrect, that could carry health consequences.

One interesting difficulty is the sparsity or impulse like nature of signals in the multivariate time series domain of the insulin pump data. This is evident in the under-performing results from the DNN model. An analogy could be made to a “needle in a haystack”. Although attempts were made to correct for this sparsity by smoothing with rolling windows and by de-noising with low pass filters, there are likely large feature engineering gains to be made with this technique.

Another challenge was that throughout many contexts, motifs appeared to be similar. This speaks to the presence of underlying wavelets or patterns that exist across all insulin pump patients.

Finally, an important limitation of the models and pipelines highlighted in this study is that real patient data was not used. Due to the nature of predicting simulation hyper parameters, it may actually not be possible to predict actual pump therapy settings like ISF, ICR, and BR in a supervised manner without a proxy feature.
5.3.2 Future Work

In order to truly test the efficacy of the models in research, a next step would be to verify the efficacy of real insulin pump data. This is an obvious next step for any model that would be used in the medical device industry, especially for predicting pump therapy settings.

Using real subjects for evaluation not only requires a proxy feature for actual values of ISF, ICR, and BR but also requires de-identification, HIPAA compliance, data cleanup and regularization. It comes with a host of complexities and risk highlighted in the ethical concerns section.

Motifs

In order to mitigate the challenge of contextual motifs looking very similar, a future goal might be to run both motif detection across a wider range of motif window sizes and well as to de-prioritize already detected motifs which appear in many contexts. This would likely encourage motif detection to find a broader context of unique motifs.

For large data sets, this comes with the cost of high computational complexity. Libraries exist to massively parallelize this process. While GPU’s were used in this research, DASK clusters or other multi-compute, multi-node architectures could be leveraged as well.

DNN

Further research can also be done to attempt to increase the performance of the time series native models like DNN explored in this study by implementing a stricter feature selection methodology to filter out noise from the contextual sub-sequences.

Another area of exploration could be to create a multi-input network and feed in different subsets of features for ISF, BR, and ICR. This could mitigate noise across the times series.

Yet another area of future interest would be to use self-attention networks to focus a DNN on certain regions of the time series (possibly post-meal). It is likely that a time series native model can perform well, and that unlocking it is a matter of the right preprocessing. In theory, there is more data present in the raw time series data than aggregation based data, due to the non-linear compression that occurs in aggregation. That said, the question remains if it is possible in practice to beat out aggregation based methods.

Ensemble

Another opportunity for further research could be building an ensemble model including both aggregation and non-aggregation based models. For this idea to work, the data pipeline would need to be able to handle both time series and aggregate data simultaneously. This would add an additional layer of complexity into the predictive pipeline which might not be desirable for real-time predictions on insulin pump devices.
5.3.3 Ethical Concerns

As with many studies in the medical device field, ethical concerns are present. Although data in this study was simulated, one of the next progressions for this study would be to use real patient data. While this data would certainly need to be de-identified, it would still be sensitive. Both permission from patients and special care is needed to handle this data appropriately.

Another area of concern is that the data used in this study was generated with a simulator meant to represent the population mean of people with type 1 diabetes in the United States. This affects how broadly the models in this study generalize to people outside of the United States.

Another concern that arises when trying to use real patient data, is that the actual values for ISF, ICR, and BR are not known. Instead, it may be difficult to find patients where close to actual values of ISF, ICR, and BR are known. Even more difficult would be to ask those patients to intentionally set their pumps to the incorrect values for the sake of this study. Including real patients would bring about numerous risks.

6 Conclusion

This research proposes novel and effective methods for predicting actual values of Insulin Sensitivity Factor (ISF), Insulin to Carbohydrate Ratio (ICR) and Basal Rate (BR) for the people with insulin pumps. This prediction is done from 30 days worth of pump data rather than existing iterative approaches, which can take up to six to eight weeks to converge on actual physiological values. This results in at least two weeks of healthier life for patients.

A Tandem Diabetes Care proprietary data simulator was leveraged to generate data spanning 1,000 multivariate time series. This data was used to train competing predictive models, and a 20% holdout subset was used for testing.

To capture unique and possibly subtle patterns which correlate with patient contexts like low ISF, or low ICR, and also patterns which correlate to pre- and post-prandial contexts, contextual consensus motif analysis was included in the pre-processing pipeline used in this study. While this method did yield features that were ultimately used in this study, time series contextual consensus motifs are an interesting area of future research. They provide a quick and interpretable snapshot of dominant patterns in a given time series.

The time series native models assessed were mainly variants of DNNs. However, DNNs were outperformed by aggregation models in terms of Mean Absolute Error. DNNs instead performed roughly similar to a No-Information Naive model. This was possibly due to signal sparsity.

Aggregation based models broadly outperformed time series native based models when predicting actual values of pump settings as a percentage of existing programmed values. Of the aggregation based models, the best performing was Ridge regression. Certain features like percent time clipping were determined to be most important when predicting ISF, ICR, and BR.
It is interesting and also fortunate that a complex time series native model was outperformed by a simpler aggregation based one. It could be that aggregation based models are better suited to solve this problem.

The goal of this study was to demonstrate techniques which can be paired with existing iterative and reinforcement learning algorithms for predicting actual patient pump settings. Hopefully, the techniques highlighted provide useful insight to future researchers, as well as improving diabetes care and quality of life for people using insulin pumps.

References


