On the Use of Consumer-grade Activity Monitoring Devices to Improve Predictions of Glycemic Variability*

Chandra Krintz[1], Rich Wolski[1] Jordan E. Pinsker[3], Stratos Dimopoulos[1], John Brevik[4], and Eyal Dassau[2]

[1] Computer Science Dept., University of California, Santa Barbara,
ckrintz@cs.ucsb.edu.edu,
http://www.cs.ucsb.edu/~racelab
[2] Chemical Engineering Dept., University of California, Santa Barbara,
[3] William Sansum Diabetes Center, Santa Barbara, CA
[4] Mathematics Dept., California State University, Long Beach

Abstract. This paper examines the use of partial least squares regression to predict glycemic variability in subjects with Type I Diabetes Mellitus using measurements from continuous glucose monitoring devices and consumer-grade activity monitoring devices. It illustrates a methodology for generating automated predictions from current and historical data and shows that activity monitoring can improve prediction accuracy substantially.

Key words: activity monitoring, diabetes, prediction, decision support

1 Introduction

Ubiquitous mobile Internet connectivity has driven the rapid development of consumer-grade "wearable" devices and monitors that collect data about the wearer, including measurements of movement (or lack thereof), exercise regimes, sleep patterns, vital physiological statistics, activities, and environment. Users employ such data to motivate active lifestyles and to provide new insights into, and decision support for, good health and well-being. Recent technological advances have driven down the cost, size, and ease of use of such devices significantly, making them accessible to a large portion of the population. Such accessibility and potential for widespread use has, in turn, spawned a variety of device options, applications, and analytics technologies.

The goal of our work is to investigate the impact of using activity monitor data to improve the health and well being of individuals with Type I Diabetes Mellitus (T1D). Diabetes is a family of chronic diseases that impacts how the human body produces and uses insulin. When food is eaten, the body digests

* This work has been supported, in part, JDRF grant number 17-2013-473, NSF grants CCF-1539586, CNS-0905237, CNS-1218808, and ACI-0751315, and NIH grant 1R01EB014877-01. We also acknowledge Wendy C. Bevier and Paige K. Bradley for the William Sansum Diabetes Center for their indispensable contributions.

and converts it to glucose (sugar). The blood glucose levels of individuals with diabetes can vary significantly. If left unchecked and unregulated, *glucose variability* and, in particular, glucose levels outside the normal range can lead to serious complications across the body's systems (vision, hearing, skin, nervous, cardiovascular, and others), which can lead to death.

For this reason, researchers have studied a wide variety of methods for measuring glycemic variability (GV) [15, 12, 1, 14, 9, 2, 5] with the goal of producing a risk index for T1D complications (c.f. Table 1 in Section 2) and informing treatment. There is as yet no consensus as to what is the most effective GV risk index, spurring further research and, almost certainly, the definition of new indices in the future.

The American Diabetes Association recommends that people with T1D engage in frequent moderate aerobic physical activity as part of their daily glucose management. However, exercise affects glucose variability and is associated with an increased risk of hypoglycemia (low glucose levels). Hypoglycemia risk is greatest during exercise, 2-3 hours following activity, and as a latent effect 12-18 hours after the activity [7, 17, 4]. Currently, people with T1D who perform physical activity must plan ahead of the event to prevent exercise related hypoglycemia. Use of temporary basal rates to reduce insulin delivery, suspension of insulin during and after the event, and carbohydrate intake can prevent immediate glucose drop. This sequence of steps prevents spontaneous activity and requires patients' compliance and significant focus and forethought to prevent adverse events. Thus if GV risk can be predicted automatically from CGM and non-invasive activity monitoring, individuals with T1D can use these predictions to inform their calorie consumption, insulin intake, and exercise decisions.

In this work, we investigate the development of an automated system for predicting next-day GV risk-index values. Our approach combines activity monitoring measurements gathered continuously from consumer-grade devices with blood glucose measurements taken by a continuous glucose monitoring (CGM) device for the same time period. Using Partial Least Squares (PLS) [16] regression, the system generates a model for predicting GV values for the day following a day when activity and CGM measurements have been gathered.

Rather than focusing on a single GV risk index, however, the methodology chooses the "best" index for a specific individual amongst a set of index choices. The "best" index in our study is the one that exhibits the most predictability (i.e. the lowest prediction error) when its automatically generated model is cross-validated.

Thus the methodology is flexible and adaptive. Each time the data is analyzed, the system may choose a new risk index as being most predictable for a given individual. As new GV indices are developed, they can be added to the suite of indices the system comprises. Similarly, if behavior or physiology changes degrade GV predictability for a given index, the method can determine which new index should replace it as being most predictable.

In this paper, we describe the algorithmic and statistical approaches we use as the basis for this system and detail their function using data gathered from a small clinical study. Our results indicate that the combination of activity data with CGM measurements can improve GV predictability, in some cases substantially.

2 Predicting Glycemic Variability

The goal of the study is to determine the extent to which activity data gathered by consumer-grade activity-monitoring devices (e.g. those manufactured by Garmin [6] or Jawbone [8]) enhance the predictability of glycemic variability (GV) [15]. GV measures and indices are the subject of much research [15, 12, 1, 14, 9, 2, 5]. Rather than choosing a single metric, our approach predicts a set of metrics for each subject to determine which in the set yields the most predictability. That is, rather than attempting to differentiate between the metrics in terms of efficacy, we assume that there is a "best" metric (in terms of predictability) for each subject. Our system automatically identifies this most predictable GV metric from a database of continuous glucose monitoring (CGM) measurements and activity measurements that are gathered on a per-subject basis.

2.1 Data Gathering

In this work, we captured the CGM and activity measurements for seven subjects with T1D during a six-week clinical study. Our study population included adults aged 18-75 years and a mix or those who are typically active and/or perform exercise as part of their weekly routine, and those who are more sedentary, to assess the validity of the algorithms in detecting exercise and determining if any false positive detections occur. Subjects were diagnosed with T1D at least 1 year or more prior to enrollment. The purpose of this is to avoid the transitory "honeymoon" phase in which beta-cells maintain significant insulin production. Subjects had an A1c value less than or equal to 9.5% at the start of the trial, as chronic hyperglycemia is indicative of problems in addition to T1D, or an inability/unwillingness to effectively participate in self-treatment.

Each subject wore both Garmin and Jawbone activity-monitoring devices simultaneously as well as a a Dexcom [3] CGM device. The CGM device records blood glucose levels every 5 minutes. However, the Garmin and Jawbone consumer activity devices do not make fine-grained measurements immediately available for download. Instead, the data gathered by each is uploaded to a proprietary service, where it is summarized (either as average or aggregate) on a 24-hour basis before it is available for download and analysis. Thus, the minimum future timeframe over which any prediction is possible using these devices as consumer goods is 24 hours. We believe that each device stores data at a finer time granularity and that if that data were made available, shorter-term predictions would be possible.

For this study, we aggregate the CGM data on a daily basis (midnight to midnight) so that it matches the time resolution of the activity data. We then predict the next day's GV measurements from each day for which both CGM and activity data are available. Note that, as in any clinical study, the data is subject to some "dropout" – periods of time when one or more measurements are missing. While the duration of the complete study is 6 weeks, only a subset of the period contains complete datasets from the CGM, Jawbone, and Garmin devices. It is from the days in the study when all three measurements are available that we make predictions. Note that only the CGM measurements need be available for the day following a day when CGM and activity data are available. Each GV measure depends only on CGM data. Thus, to validate the predictions, we identify 24-hour periods in the data for each subject in which

- CGM measurements, Garmin summaries, and Jawbone summaries are available, and
- CGM measurements are available for the following 24 hour period.

The system then predicts GV metrics computed for the following day from the GV metrics and activity data for each day in which the data is available. Table 1 summarizes the GV metrics we consider for each subject.

The activity monitors offer several different measurements of activity, including number of steps taken (pedometer), estimated calories burned, maximum period of activity, and number of sleep hours. Studying these measurements revealed that some of them (e.g., calories) are computed directly from the others arithmetically. Also a cursor inspection of some of the values reveals that they can be unreliable (for example 23 hours of sleep in a 24-hour period). Table 2 shows the activity measures we believed to be suitable for this study. Note that we were only able to get a pedometer reading from the Garmin device consistently during the study. We include both pedometer measurements because in the case where both Jawbone and Garmin pedometer measurements are available, they differ substantially enough to warrant their inclusion as separate measurements. Also, we made no attempt to "sanity-check" the data to determine whether it is valid. For example, the *acttime* and *inacttime* should sum to 24 hours for each day, and they do not. Rather, as long as each measurement seemed "feasible" in isolation, we included it.

3 Prediction Methodology

The basis of the prediction methodology we explore is least-squares regression [16]. Our goal is to determine a linear model with the various regressors (GV and activity measurements for a 24-hour period) that best predicts a specific GV metric for the following day. That is, for a $m \times n$ matrix A consisting of m GV and activity values on each of n days, and a vector $\mathbf{y}$ consisting of n GV values for each succeeding day, we will find the approximate solution $\hat{\mathbf{x}}$ to the (typically overdetermined and unsolvable) equation

$$A\mathbf{x} = \mathbf{y} \tag{1}$$

that is best in the sense that $\|A\hat{\mathbf{x}} - \mathbf{y}\| \leq \|A\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x} \in \mathbb{R}^m$.

Metric	Formula	Description
bgavg	$\frac{\sum_{i=0}^{n} BG_i}{n}$	Average daily BG[12]
bgmin	$\min_{i=0}^{n} BG_i$	Min daily BG[12]
bgmax	$\max_{i=0}^{n} BG_i$	Max daily BG[12]
drop	$\max_{i=0}^{k} drop(BG)$	Max k daily continuous BG drops
spike	$\max_{i=0}^{k} drop(BG)$	Max k daily continuous BG increases
stdev	$\frac{\sum_{i=0}^{n} \lvert BG_i - bgavg \rvert}{n-1}$	Standard deviation of BG[1, 2, 12, 15, 14]
jindx	$0.001 * (bgavg + stdev)^2$	J-Index: Glycemic control quality metric[15]
covar	$100 * (bgavg * stdev)$	Coefficient of Variation[1]
mag	$\frac{\sum_{i=0}^{n-1} \lvert BG_i - BG_{i+1} \rvert}{24}$	Mean absolute BG difference[15]
range	$bgmax - bgmin$	Min-max BG difference for day[1]
med	$median(BG_1^n)$	Median daily BG[2, 1]
iqr	$Q3(BG_1^n) - Q1(BG_1^n)$	Interquartile daily range (IQR)[1]
low	$\sum duration(< 70)$	Time spent below 70 BG (hypo)[12]
high	$\sum duration(> 180)$	Time spent above 180 BG (hyper)[12]
tg	$low + high$	Time spent outside of BG range (70-180)[12]
mage	$\frac{\sum_{i=0}^{k} (PND_k > stdev)}{k}$	Mean amplitude of glycemic[2, 12, 1] excursions[15, 14] PND = BG peak-nadir differences
mavg	$\frac{\sum_{i=0}^{n} \lvert 10 * \log BG_i / 80 \rvert^3}{n}$	M-value: Glycemic control quality metric[1, 15]
mravg	$\frac{\sum_{i=0}^{n} 1000 * \lvert \log BG_i / 100 \rvert}{n}$	MR-value: Weighted average of BG values[2] (control quality metric)
grade	$\frac{\sum_{i=0}^{n} 425 * (\log(\log(BG_i * 18) + 0.16))^2}{n}$	Glycemic risk assessment diabetes equation[15, 14]
modd	$\frac{\sum_{k=d2, t=s}^{dp, e} \lvert BG_{k,t} - BG_{k-1,t} \rvert}{p-1}$ s=midnight, e=11:59:59	Mean of daily differences[2, 14] $p - 1$ consecutive days starting day 2
adrr	$\frac{1}{M} * \sum_{i=1}^{M} (LR^i + HR^i)$ $LR^i = \max(rl_{dayi}(...))$ $HR^i = \max(rh_{dayi}(...))$ for days $i = 1 - M$ and day_i BG values $j = 1 - N$	Average Daily Risk Range[14, 2] $TBG_j = 1.509 * ((\ln(BG_j)^{1.084}) - 5.381)$ $rl(BG_i) = 10 * TBG_i^2$ if $TBG_i < 0$,else 0 $rh(BG_i) = 10 * TBG_i^2$ if $TBG_i > 0$,else 0
lbgi	ADRR only for LR^i only	[2, 14]
hbgi	ADRR only for HR^i only	[2, 14]
conga N	$\sqrt{\frac{\sum_1^k (DT - AD)^2}{k-1}}$ $DT = BG_t - BG_{t-m}$	Continuous overall net glycemic action[15, 2] $AD = \frac{\sum_1^k DT}{k}$, $m = 60 * N$, $N = 1, 2, 4, 6$[1, 12, 14]

Table 1. Measures of Glycemic Variability, formulas, descriptions, and citations. We define *daily* as midnight to 11:59:59PM of a given day. BG is blood glucose in mg/dL.

However, direct application of ordinary least-squares (OLS) regression [10] to the problem of predicting next-day GV values is problematic in our setting. While the study spans a 6-week period, each subject did not wear all of the devices (CGM and activity monitoring) each day. Indeed the number of monitored days in the study varies between 28 for subject 001 to 39 for subject 004. With 27 GV metrics and 6 activity measures, the total number of regressors (in this case

Measurement	Description	Vendor
accttime	total time active in a day	Jawbone
inacttime	total time inactive in a day	Jawbone
maxact	max period of continuous activity in a day	Jawbone
maxidle	max period of continuous inactivity in a day	Jawbone
jsteps	steps taken in a day	Jawbone
gsteps	steps taken in a day	Garmin

Table 2. Activity measurements and device vendor. Units of time are seconds.

the m GV and activity measurements represented in the A matrix) is close to the number of measured GV values (one computed for each of the n days during which a subject wore the CGM and activity monitoring devices). Additionally, many of the predictors are correlated with one another, for example *acttime* and *maxact* from Table 2, and these dependencies create instabilities in solutions of linear problems such as the one above. For these reasons, we seek a model that is parsimonious, using a small number of predictors that nevertheless capture most of the predictive power of the entirety.

Partial Least Squares regression (PLS) [16] is a linear regression technique that attempts to identify a smaller number latent factors in the regressor set that best predict the value of the target variable (next-day GV value in our case). It does so by transforming the regressors and the target variable so that the multi-dimensional variance in the transformed regressor space best explains the variance in the transformed target. PLS is often a better choice of technique than Principal Components Regression [11] or Ridge Regression [13] (two related methods) when the goal is to minimize prediction error and the explanatory value of any one regressor is not required.

In this work, we are interested in determining whether an automated technique based on PLS is feasible. To do so, we must define a method for determining the number of latent factors to use in instance of PLS. We use a form of cross-validation (see Subsection 3.3) to identify the set of factors that results in the minimum prediction error. Specifically, we consider latent factor counts (which we henceforth term "component counts") from 1 to 10 and compute the cross-validation prediction error for each target GV value associated with each count. The component count that corresponds to the smallest prediction error is then selected as the component count to use.

Thus, the PLS method first identifies the component count to use. It then generates a model in the transformed space using this component count constructs a linear model in the untransformed space of the regressors and target variable. In our setting, this linear model takes the values of the regressor variables gathered on a specific day and predicts a specific GV value for the next day.

3.1 Categorization

In examining the data, we observed that several of the subjects experienced "high" and "low" days, particularly with respect to activity. Further, it seemed (by inspection) that the relationship between regressors and predicted GV for the next day differed depending on whether activity levels were "high" or "low".

To test whether predictions are improved by categorization, we also include the possibility of running separate regressions (each using PLS) using only "high" days or "low" days as categorized by a specific regressor. For example, it may be that the `mage` [15, 14] index is more predictable after a day of high activity than it is generally or after a day of low activity. Further, the best activity measure (as reported by the activity device) used to categorize "high" and "low" effectively might vary by subject.

For the purposes of categorization we use one-dimensional k-means clustering. That is, a single metric is clustered into two categories, high and low. The regressors associated with the values in the high category as well as the next-day's target GV metric are extracted. We then use PLS on the extraction to compute predictions of the GV metric on days following a day of high activity. Similarly, the method can automatically extract regressors corresponding to low days so that they may be used separately to make predictions via PLS.

3.2 The Algorithm

As described in Section 2, the full regressor set consists of 27 GV metrics and 6 activity measures (c.f. Tables 1 and 2 respectively). Each GV metric can be a prediction target, each regressor can be used to categorize the time epoch into "high" and "low", and there are 7 subjects in the study. The full regression analysis algorithm for data generated by the study is shown in algorithm 1. The output of the algorithm is a set of linear models (produced on lines 4, 9, and 10 respectively). Each model predicts a different GV metric for a specific subject using either *all* of the regressors (line 4), or based on a categorization into high and low days (lines 9 and 10). The subscripts index each model. For example, the model predicting the *grade* metric for subject 1 using all of the regressors is indexed as $LinearModel_{1,grade,all}$. Similarly, the linear model predicting the *modd* metric for subject 5 using *high* days only and *acctime* to split regressors into high and low sets is denoted $LinearModel_{5,modd,acttime,high}$.

Algorithm 1 Full Regression Analysis Algorithm

```
 1: for each subject $S$ do
 2:     $RegressorSet \leftarrow$ regressors from days with valid data for $S$
 3:     for $GV$ in set of GV metrics do
 4:         $LinearModel_{S,GV,all} \leftarrow PLS(RegressorSet, GV)$
 5:         for $R$ in $RegressorSet$ do
 6:             divide $R$ into two clusters using k-means
 7:             $HighRegressors \leftarrow$ regressors from days in high cluster
 8:             $LowRegressors \leftarrow$ regressors from days in low cluster
 9:             $LinearModel_{S,GV,R,high} \leftarrow PLS(HighRegressors, GV)$
10:             $LinearModel_{S,GV,R,low} \leftarrow PLS(LowRegressors, GV)$
```

3.3 Evaluating Model Fitness and Predictive Power

The system attempts to identify automatically which model makes the most accurate predictions for each subject, and the specific GV metric that is best predicted by that model. To do so, it uses the regression's R^2 measure to determine the degree to which each linear model explains the variation about the mean value for a specific GV metric. An R^2 value close to 1.0 indicates that almost all of the difference between the observed vales of a GV metric and the values predicted by the model are due to random variation. Alternatively, a value close to 0.0 indicates that the model does not account for much of the variance in the distribution of GV values. Thus values closer to 1.0 indicate a better "fit" of the model.

For each subject, we concentrate on linear models where the R^2 statistic is greater than or equal to 0.85. The cutoff for deciding whether a fit is good ($R^2 >= 0.85$) is somewhat arbitrary, indicating that each model under consideration explains at least some of the variance.

To measure predictive accuracy, we compute the root mean square prediction error for each observed value of the specific GV metric under consideration using an "all-but-one" validation strategy (also termed "cross-validation" in some settings). That is, for each linear model for which the R^2 value is greater than or equal to 0.85 (indicating a linear relationship) for a specific GV metric, we

- remove a test value from the set of values gathered for the GV metric
- remove the set of regressors for the day before this test value was recorded
- predict the test value using the PLS algorithm with the remaining regressors and the remaining GV values, and
- record the prediction error as the difference between the predicted GV value and the test GV value

We repeat this procedure making each value in the set of GV values the test value. That is, a prediction error is generated for each value in the set of GV values when the the other values are used to "train" the system. The square root of the mean of the squared prediction errors is the overall cross-validate Root Mean Squared Prediction Error (RMSPE). Smaller RMSPE values indicate more accurate predictions.

Figure 1 depicts the functional decomposition of the automated system. To make a prediction for each subject, the system requires access to a database of previous CGM measurement history and a database of activity measurements. Note that all of the GV metrics described Table 1 can be computed from a single set of CGM measurements. Thus the first step is to convert the history of CGM measurements into a history of GV measurements for each GV metric configured into the system. The GV histories are combined with activity history in Algorithm 1 to create a set of linear models. The system uses a selection mechanism (in this study, it selects the model with the minimum cross-validated RMSPE) to pick the best linear model for the most predictable GV. The current CGM data are converted to this GV metric and combined with the current activity data to predict the next day's GV value. Once the next day's CGM

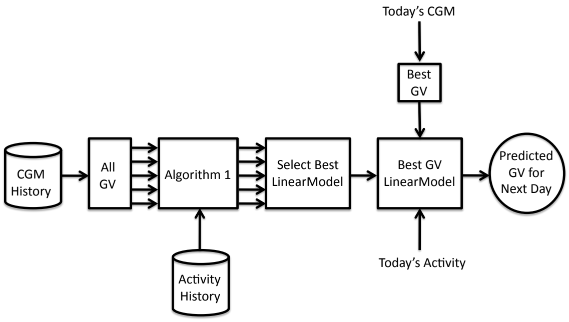

Fig. 1. Functional decomposition of automated prediction methodology

Subject	Predicted GV	R^2	Components	Sample Size
001	**covar**	0.88	10	26
002	**mravg**	0.52	10	35
003	**modd**	0.69	10	30
004	**cnga6conga**	0.72	10	37
005	**spike**	0.56	10	33
006	**mravg**	0.71	10	36
007	**addrDRR**	0.80	10	28

Table 3. Baseline best regression results by subject using GV values only.

and activity data are available, they can be added to their respective historical databases and the process repeated.

4 Results

We begin with an examination of the effectiveness of the regression technique described in Section 3 using GV values only. In this "baseline" analysis, the regressors consist of the 27 GV metrics shown in Table 1. For each subject, we compute the linear model that predicts each of the 27 metrics for the succeeding day and show the model that generates the largest R^2 value for each subject in Table 3. In the table, column 1 gives the subject identifier, column 2 shows the specific GV metric for which the largest R^2 value was generated for that subject, column 3 shows the R^2 value for that metric, column 4 shows the number if principal components, and column 5 shows the sample size (i.e. the number of days n for which there is valid GV data for that subject in the study).

Note that using GV values alone produces a regression for only one subject (subject 001) with an R^2 greater than 0.85. Using an R^2 cut-off of 0.85 as described in Section 3 the results in Table 3 do not indicate a strong linear relationship between current and next day GV measures.

In Table 4 we add to the regressor set the 6 activity measures described previously in Table 2 and repeat the analysis. While only one of the R^2 values reach the 0.85 threshold, comparing Tables 3 and 4 shows that the addition of activity data to GV metric does improve regression fit. Subject 001's R^2 value

Subject	Predicted GV	R^2	Components	Sample Size
001	**covar**	0.84	10	26
002	**mag**	0.76	10	35
003	**bgavg**	0.76	10	30
004	**cnga6conga**	0.72	10	37
005	**spike**	0.65	10	33
006	**low**	0.73	10	36
007	**addrDRR**	0.85	10	28

Table 4. Baseline and activity measures in the regressor set, best regression results for GV by subject.

Subject	Category Discriminant	Predicted GV	R^2	RMSPE % Improvement	Components	Sample Size
001	sddev	**bgmin**	0.99	40%	10	12
002	gsteps	**modd**	0.99	36%	10	16
003	sddev	**bgmax**	0.95	40%	6	11
004	hbgi	**high**	0.99	26%	10	11
005	covar	**covar**	0.85	34%	7	12
006	grade	**high**	0.99	50%	10	11
007	inacttime	**spike**	0.99	41%	10	13

Table 5. Most improved regression predictions of next-day GV by subject for days categorized as "high." Baseline and activity measures are included in the regressor set, only regressions with R^2 greater than or equal to 0.85 are included.

drops from 0.88 to 0.84 (which is almost above the threshold) while the best R^2 value from each of the other subjects increases with the addition of activity data to the regression.

Also notice that the GV metric exhibiting the best R^2 value changes for three of the subjects (002, 003, and 006) when activity data is introduced. From the results shown in Tables 3 and 4 we cannot conclude (based on R^2 value) that PLS is an effective predictive technique for all subjects and GV metrics. However it does appear that the addition of activity data from consumer-grade activity monitors improves the linear fit generated by PLS with respect to next-day GV when the data is uncategorized.

4.1 Prediction After Categorization

Separating the per-subject GV and activity measurement data into high and low categories improves regression performance in terms of R^2 value. In Table 5 we show the effect of including activity data on predictability of GV per subject. Each row of the table shows the GV metric (in column 3 in boldfaced type) for which activity data results in the greatest improvement (greatest reduction) of prediction error, as measured by RMSPE (c.f. Subsection 3.3). Column 2 of the table indicates which GV or activity metric value is best used to categorize the data into "high" and "low" groups of days. Column 4 shows the R^2 value for the PLS, column 5 shows the percentage improvement in RMSPE, column 6 shows the number of principal components, and column 7 indicates the sample size.

Subject	Category Discriminant	Predicted GV	R^2	RMSPE % Improvement	Components	Sample Size
001	drop	**drop**	0.95	50%	6	13
002	hbgi	**conga 1**	0.99	55%	10	13
003	conga 4	**mag**	0.99	49%	7	11
004	tg	**grade**	0.95	41%	10	14
005	range	**grade**	0.99	54%	10	15
006	iqr	**drop**	0.99	59%	9	16
007	tg	**sddev**	0.99	49%	9	14

Table 6. Most improved regression predictions of next-day GV by subject for days categorized as "low." Baseline and activity measures are included in the regressor set, only regressions with R^2 greater than or equal to 0.85 are included.

Note that we calculate the percentage improvement over *the best* prediction of the GV metric (in terms of RMSPE) that can be made for that individual when activity data is not considered. That is, splitting the data according to the metric shown in column 2 might result in the lowest RMSPE when activity data is considered, but a different split might result in a lower RMSPE when just CGM data is considered. When calculating percentage improvement, we use the lowest RMSPE for the CGM-only comparison across all categorizations and not just the one that minimizes RMSPE when activity data is in the regressor set.

For example, row 1 of Table 5 shows, for subject 001, that activity data improves the RMSPE for the **bgmin** GV metric when the **sddev** metric is used to divide days into those with a high **sddev** score and those with a low **sddev** score. The R^2 value for the PLS on the high data is 0.99 and the improvement in RMSPE for subject 001's **bgmin** GV measure is 40%. There were 10 components selected for the PLS and 12 days qualified as "high" days in terms of **sddev** (sample size is 12).

Table 6 shows the most improved RMSPE GV metrics for days automatically categorized as "low." From both Table 5 and Table 6 it appears that activity data improves PLS predictability when the data is bifurcated for each patient into two categories: "high" and "low". The discriminant that results in the greatest improvement (column 2 in both tables) varies by subject, as does the GV metric that experiences the greatest improvement in predictability.

4.2 Conclusions

The results seem to indicate that activity monitoring via consumer-grade "wearable" devices does improve next-day GV predictability, via PLS, in some measure. While the R^2 values in Table 5 and Table 6 are likely overstating the linear nature of the model because the number of PLS components is close to the sample size, the improvement metric is based on cross-validated RMSPE. Further, because each specific split and subsequent regression with activity data in the regressor set is compared to the best cross-validated RMSPE without activity data, we believe that these results show that activity data improves predictability for some of the metrics captured in the study. Also, the specific metric and data categorization that works "best" varies by subject (and indeed may vary by time although our study does not explore time variation). This observation argues for a fully-automated system that can recomputes future GV for each subject when new data becomes available (e.g. every day when the activity devices summarize user measurements).

References

1. F. Cameron, P. Baghurst, and D. Rodbard. Assessing glycemic variation: why, when, and how? *J Pediatric Endocr Reviews*, 7(3), 2010.
2. D. Czerwoniuk, W. Fendler, L. Walenciak, and W. Mylnarski. GlyCulator: A Glycemic Variability Calculation Tool for Continuous Glucose Monitoring Data. *J Diabetes Sci Technol.*, 5(2), 2011.
3. Dexcom. `https://www.dexcom.com/`. [Online; accessed 25-March-2015].
4. C. Ellingsen, E. Dassau, HC. Zisser, B. Grosman, MW. Percival, L. Janovic, and FJ. Doyle III. Safety constraints in an artificial pancreatic -cell: an implementation of model predictive control with insulin-on-board. *J Diabetes Sci Technol.*, 3(3), 2009.
5. LS. Farhy, EA. Ortiz, BP. Kovatchev, AG. Mora, SE. Wolf, and CE. Wade. Average Daily Risk Range as a Measure of Glycemic Risk Is Associated with Mortality in the Intensive Care Unit: A Retrospective Study in a Burn Intensive Care Unit. *J Diabetes Sci Technol.*, 5(5), 2011.
6. Garmin. `https://www.garmin.com/`. [Online; accessed 25-March-2015].
7. R. Gondhalekar, E. Dassau, HC. Zisser, and FJ. Doyle III. Periodic-Zone Model Predictive Control for Diurnal Closed-Loop Operation of an Artificial Pancreas. *J Diabetes Sci Technol.*, 7(3), 2013.
8. Jawbone. `https://www.jawbone.com/`. [Online; accessed 25-March-2015].
9. B. Kovatchev, E. Otto, D. Cox, L. Gonder-Frederic, and W. Clarke. Evaluation of New Measures of Glucose Variability in Diabetes. *J Diabetes Care*, 29(11), 2006.
10. Ordinary Least Squares. `https://en.wikipedia.org/wiki/Ordinary_least_squares`. [Online; accessed 25-March-2015].
11. Principle Components Regression. `http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Principal_Components_Regression.pdf`. [Online; accessed 25-March-2015].
12. RA. Rawlings, H. Shi, LH. Yuan, W. Brehm, R. Pop-Busui, and PW. Nelson. Translating Glucose Variability Metrics into the Clinic via Continuous Glucose Monitoring: A Graphical User Interface for Diabetes Evaluation. *J Diabetes Technol and Ther.*, 13(12), 2011.
13. Ridge Regression. `http://ncss.wpengine.netdna-cdn.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Ridge_Regression.pdf`. [Online; accessed 25-March-2015].
14. D. Rodbard. Interpretation of continuous glucose monitoring data: glycemic variability and quality of glycemic control. *J Diabetes Technol and Ther.*, 11(1), 2009.
15. FJ. Service. Glucose Variability. *J Diabetes*, 62(5), 2013.
16. Randall D Tobias et al. An introduction to partial least squares regression. In *Proc. Ann. SAS Users Group Int. Conf., 20th, Orlando, FL*, pages 2–5. Citeseer, 1995.
17. K. van Heusden, E. Dassau, HC. Zisser, DE Seborg, and FJ. Doyle III. Control-relevant models for glucose control using a priori patient characteristics. *IEEE Trans Biomed Eng.*, 59(7), 2012.